Degree Project in ?

Second cycle, 30 credits

Reducing read and write latency using Rust in an offline feature store

Developing delta-rs to support HopsFS and reducing read and write latency on the Hopsworks offline feature store

GIOVANNI MANFREDI

Reducing read and write latency using Rust in an offline feature store

Developing delta-rs to support HopsFS and reducing read and write latency on the Hopsworks offline feature store

GIOVANNI MANFREDI

Master's Programme, ICT Innovation, 120 credits
Date: October 18, 2024

Supervisors: Sina Sheikholeslami, Fabian Schmidt, Salman Niazi
Examiner: Vladimir Vlassov
School of Electrical Engineering and Computer Science
Host company: Hopsworks AB
Swedish title: Detta är den svenska översättningen av titeln
Swedish subtitle: Detta är den svenska översättningen av undertiteln

Abstract

The need of building Machine Learning (ML) models based on ever increasing amounts of data brought new challenges to data management systems. Feature Stores have emerged as an effective solution to enable feature reuse, while organizing data transformations and ensuring consistency between feature engineering, model training and model inference. Recent publications demonstrate that the Hopsworks Feature Store exhibits superior performance metrics in both training and online inference query workloads compared to existing cloud-based alternatives. In this system, the latency to perform a write operation in this system a least of one or more minutes even for small quantities of data (1 GB or less). This limitation is believed to a Spark-specific limitation, which the system uses to write data on the offline feature store. This hypothesis was already confirmed in the case of read latency, where opting for a Spark-alternative, namely an Arrow Flight and DuckDB server, improved performance considerably. A promising approach appears to be adopting a new data management solution, namely Delta Lake, and accessing it using a Rust library called delta-rs. This thesis investigates the possibility of reducing the read and write latency in the offline feature store by expanding the delta-rs library to support the Hopsworks Feature Store file system called HopsFS, and comparatively evaluating the performance of the legacy and newly implemented system. After a first iterative system implementation phase based on fixed requirements, the system was evaluated by performing and measuring read and write operations in a four different CPU configurations (increasing the number of CPU cores up to eight). Experiments were performed fifty times to estimate a confidence interval allowing a comparative evaluation of the systems. Results confirmed the superior performance of the delta-rs library over the Spark system in all write operations with a tenfold reduction in the latency. Delta-rs also surpassed the Spark-alternative in read operations with a tenfold reduction in the latency in all but the experiment with the largest table (60M rows), where the improvement is of a smaller factor. These findings encourage future research investigating Spark-alternative when optimizing performance in small scale (1 GB - 100 GB) data management systems.

Keywords

Canvas Learning Management System, Docker containers, Performance tuning

Sammanfattning

Här ska jag skriva ett abstract som är på ca 250 och 350 ord (1/2 A4-sida) med följande komponenter:

- Vad är ämnesområdet? (valfritt) Presenterar ämnesområdet för projektet.
- Kort problemformulering
- Varför var detta problem värt en kandidat-/masteruppsats? (*i.e.*, varför är problemet både betydande och av en lämplig svårighetsgrad för ett kandidat-/masteruppsats-projekt? Varför har ingen annan löst det än?)
- Hur löste du problemet? Vad var din metod/insikt?
- Resultat/slutsatser/konsekvenser/påverkan: Vilka är dina viktigaste resultat/
slutsatser? Vad kommer andra att göra baserat på dina resultat? Vad kan göras nu när du är klar - som inte kunde göras innan ditt examensarbete var klart?

Nyckelord

Första nyckelordet, Andra nyckelordet, Tredje nyckelordet, Fjärde nyckelordet

Sommario

Qui scriverò un abstract di circa 250 e 350 parole (1/2 pagina A4) con i seguenti elementi:

- Qual è l'area tematica? (opzionale) Introduce l'area tematica del progetto.
- Breve esposizione del problema
- Perché questo problema meritava un progetto di tesi di laurea/master? (Perché il problema è significativo e di un grado di difficoltà adeguato per un progetto di tesi di laurea/master? Perché nessun altro l'ha ancora risolto?)
- Come avete risolto il problema? Qual è stato il vostro metodo/intuizione?
- Risultati/Conclusioni/Conseguenze/Impatto: Quali sono i vostri risultati chiave/ conclusioni? Cosa faranno gli altri sulla base dei vostri risultati? Cosa si può fare ora che avete finito - che non si poteva fare prima che il vostro progetto di tesi fosse completato?

parole chiave

Prima parola chiave, Seconda parola chiave, Terza parola chiave, Quarta parola chiave

Acknowledgments

I would like to thank xxxx for having yyyy.

Stockholm, June 2024
Giovanni Manfredi

Contents

List of Figures

List of Tables

Listings

List of acronyms and abbreviations

AB	*Aktiebolag*, tr. Limited company
ACID	Atomicity, Consistency, Isolation and Durability
AI	Artificial Intelligence
API	Application Programming Interface
AWS	Amazon Web Services
BI	Business Intelligence
BPMN	Business Process Model and Notation
CIDR	Conference on Innovative Data Systems Research
CoC	Conquer of Completion
CPU	Central Processing Unit
CRUD	Create Read Update Delete
D	Deliverable
DBMS	Data Base Management System
DFS	Distributed File System
ELT	Extract Load Transform
ETL	Extract Transform Load
G	Goal
GCS	Google Cloud Storage
GPU	Graphical Processing Unit
HDD	Hard Disk Drive
HDFS	Hadoop Distributed File System
HopsFS	Hopsworks' HDFS distribution
IN	Industrial Need
JVM	Java Virtual Machine
LocalFS	Local File System
ML	Machine Learning

MLOps	Machine Learning Operations
OLAP	On-Line Analytical Processing
OS	Operating System
PA	Project Assumption
PC	Personal Computer
RAM	Random Access Memory
RDD	Resilient Distributed Dataset
RPC	Remote Procedural Call
RQ	Research Question
SDG	Sustainable Development Goal
SF	Scale Factor
SSD	Solid State Drive
SSH	Secure Shell protocol
TLS	Trasport Layer Security
TPC	Transaction Processing Performance Council
VM	Virtual Machine

Chapter 1

Introduction

Data lakehouse systems are increasingly becoming the primary choice for running analytics in large companies with over 1000 employees [1]. The data lakehouse architecture [2] is preferred over old paradigms, i.e., data warehouses and data lakes, as it builds upon the advantages of both systems, having the scalability properties of data lakes while preserving the Atomicity, Consistency, Isolation and Durability (ACID) properties typical of data warehouses [2]. Additionally, data lakehouse systems include partitioning, which reduces query complexity significantly and provides "time travel" capabilities, enabling users to access different versions of data, versioned over time [3].

Three main implementations of this paradigm emerged over time [4]:

1. **Apache Hudi**: first introduced by Uber, now primarily backed by Uber, Tencent, Alibaba, and Bytedance [5].

2. **Apache Iceberg**: first introduced by Netflix and now primarily backed by Netflix, Apple, and Tencent [6].

3. **Delta Lake**: first introduced by Databricks and now primarily backed by Databricks and Microsoft [7].

While large communities support all three projects, Delta Lake is acknowledged as the de-facto data lakehouse solution [4]. This is mainly due to Databricks, which first promoted this new architecture over data lakes among their clients around 2020 [7].

As a data query and processing engine, Delta Lake is typically used with Apache Spark [8]. This approach is effective when processing large quantities of data (1 TB or more) in the cloud, but whether this approach is effective on small quantities of data (100 GB or less) remains to be investigated [9].

DuckDB [10], a Data Base Management System (DBMS) and Polars [11], a DataFrame library, highlighted the limitations of Apache Spark. When the data volume is small (between 1 GB and 100 GB) and the architecture is processing data locally, an Apache Spark cluster performs worse than alternatives. This ultimately increases costs and computation time [12, 13].

Another aspect to remember is that thanks to its ease of use and high abstraction level, Python has become the most used programming language in the data science space [14]. Python is currently the most popular general-purpose programming language [15, 16], and it is by far the most used language for Machine Learning (ML) and Artificial Intelligence (AI) applications [17]; this is mainly thanks to its strong abstraction capabilities and accessibility. This trend can also be observed by looking at the most popular libraries among developers, where two Python libraries make the podium: NumPy and Pandas [16]. In this scenario, using a Python client for Delta Lake would be beneficial as developers would not have to resort to Apache Spark and its Python Application Programming Interface (API) (PySpark). This approach with small-scale (between 1 GB and 100 GB) use cases would improve performance significantly.

This native Python access for Delta Lake directly benefits Hopsworks *Aktiebolag*, tr. Limited company (AB), the host company of this master thesis. Hopsworks AB develops a homonymous Feature Store for ML, a centralized, collaborative data platform that enables the storage and access of reusable features [18]. This architecture also supports point-in-time correct datasets from historical feature data [19].

This presented project aims to reduce the latency (seconds) and thus increase the data throughput (rows/second) for reading and writing on Delta Lake tables that act as an offline feature store in Hopsworks. Currently, the writing pipeline is Apache Spark-based and the fundamental hypothesis of the project is that a faster non-Apache Spark alternative is possible. If successful, Hopsworks AB will consider incorporating this system implementation into the open-source Hopsworks Feature Store, significantly enhancing the experience for Python users working with smaller datasets (ranging from 1 GB to 100 GB). More generally, this work will outline the possibility of Apache Spark alternatives in small-scale use cases (between 1 GB and 100 GB).

1.1 Background

A comprehensive understanding of this project is based on three key aspects: the development of the Lakehouse architecture, the significance and workflows of Apache Spark, and the emergence of Python as a dominant language.

Data lakehouse is a term coined by Databricks in 2020 [20] to define a new design standard that was emerging in the industry that combined the capability of data lakes in storing and managing unstructured data with the ACID properties typical of data warehouses. Data warehouses became a dominant standard in the '90s and early 2000s [21], enabling companies to generate Business Intelligence (BI) insights, managing different structured data sources. The problems related to this architecture were highlighted in the 2010s when the need to manage large quantities of unstructured data rose [22]. So data lakes became the pool where all data could be stored, on top of which a more complex architecture could be built, consisting of data warehouses for BI and ML pipelines. This architecture, while more suitable for unstructured data, introduces many complexities and costs, related to the need of having replicated data (data lake and data warehouse), and several Extract Load Transform (ELT) and Extract Transform Load (ETL) computations. Data lakehouse systems solved the problems of data lakes by implementing data management and performance features on top of open data formats such as Parquet [23]. Three key technologies enabled this paradigm: (i) a metadata layer for data lakes, tracking which files are part of different tables, (ii) a new query engine design, providing optimizations such as RAM/SSD caching, and (iii) an accessible API access for ML and AI applications. This architecture design was first open-sourced with Apache Hudi in 2017 [5] and then Delta Lake in 2020 [7].

Spark is a distributed computing framework used to support large-scale data-intensive applications [24]. Developed as an evolution of the MapReduce paradigm, Spark has become the de-facto standard for big data processing due to its superior performance and versatility. Spark significantly improved its performance compared to its predecessor, i.e., Hadoop MapReduce (10 times better in its first iteration) [24] thanks to its use of in-memory processing. This means that Spark avoids going back and forth between storage disks to store the computation results. Spark, which is open-sourced under the Apache foundation as Apache Spark [25] (now referred to simply as Spark), has seen widespread success and adoption in various applications, becoming the de-facto data-intensive computing platform for the distributed computing world. While Spark is often used as a comprehensive solution [8], different solutions might be better suited for a specific scenario. An example of this is the case of Apache Flink [26], designed for real-time data streams, which prevails over Spark where low latency real-time analytics are required. Similarly, Spark might not be the best tool for lower-scale applications where

Spark's high-scaling capabilities may not be required. This is the case of DuckDB [10] and Polars [11], that by focusing on low scale (10GB-100GB) provide a fast On-Line Analytical Processing (OLAP) embedded database and DataFrame management system respectively offering an overall faster computation compared to starting a Spark cluster for to perform the same operations. This shows the possibility for improvements and new applications that substitute the current Spark-based systems in specific applications such as real-time data streaming or small-scale computation. In this project, the latter application is going to be explored.

Python can be considered the primary programming language among data scientists [27]. Many first adopted Python thanks to its focus on ease of use, high abstraction level, and readability. This helped create a fast-growing community behind the project, which led to the development of many libraries and APIs. So now, more than 30 years after its creation, it has become the de-facto standard for data science thanks to many daily used Python libraries such as TensorFlow, NumPy, SciPy, Pandas, PyTorch, Keras and many others. Python is also considered to be the most popular programming language, according to the number of results by search query (+*"<language> programming"*) in 25 different search engines [28]. This is computed yearly in the TIOBE Index [15]. The 2024 rankings reveal that Python holds a rating of 15.16%, followed by C at 10.97%. The index also highlights trends from recent years, clearly illustrating Python's rise over traditionally popular languages like C and Java, both of which Python surpassed between 2021 and 2022. This underlines the importance of providing Python APIs, particularly for programmers and data scientists, to enhance engagement and expand the capabilities of a framework.

1.2 Problem

The Hopsworks Feature Store [18] first used Apache Hudi for their Offline Feature Store, as it was the first open-sourced data lakehouse in 2017. Recently, Hopsworks AB added support for using Delta Lake as an offline feature store, following its clients' requests. Spark is used as a query engine in the system, i.e., executes the query (read, write, or delete) on the Offline Feature Store. Running the system showed that even a write operation on a small dataset, consisting of 1 GB of data or less, takes one or more minutes to complete.

This hurts Hopsworks' typical use case, which sits between tests on small quantities of data (scale between 1-10 GBs) and production scenarios on a larger scale but still relatively small (scale between 10-100 GBs).

This research's underlying hypothesis is that this slow transaction time is a Spark-specific issue. This has led Hopsworks to adopt Spark alternatives [9] for reading in their Apache Hudi system. Delta Lake supports Spark alternatives for accessing and querying the data, and of particular interest is the delta-rs library [29] that enables Python access to Delta Lake tables without using Spark. However, the delta-rs [29] does not support Hadoop Distributed File System (HDFS), and consequently Hopsworks' HDFS distribution (HopsFS) [30].

1.2.1 Research Question

This research project has the ultimate objective to evaluate and compare the performance of the current Spark system that operates on Apache Hudi to a Rust system that uses delta-rs library [29] operates on Delta Lake, using HopsFS [30]. To achieve this, support for HDFS (and thus also HopsFS) must be added to the delta-rs library [29] so that it can be compatible with the Hopsworks system. Thus, the project addresses the following two Research Questions (RQs):

RQ1: How can we add support for HDFS and HopsFS to the delta-rs library?

RQ2: What is the difference in latency and throughput between the current legacy system (Spark-based in writing) reading and writing to Apache Hudi compared to a delta-rs library-based reading and writing to Delta Lake, in HopsFS?

1.3 Purpose

This thesis project aims to contribute to reducing the read and write latency (seconds) and thus increasing the data throughput (rows/second) for operations on the Hopsworks Offline Feature Store. This study will compare the performance of the current legacy pipeline, which is Spark-based for writing, with the delta-rs pipeline on a small scale by evaluating differences in read latency (seconds), write latency (seconds), and computed throughput (rows/second). As a prospect for future work, if delta-rs is proven to be a more performant alternative (in terms of latency and data throughput), Hopsworks AB will consider integrating this pipeline into their application.

Overall implications for this thesis work are much broader considering Spark's popularity in the open source community (more than 2800 contributors during its lifetime [31]). Choosing delta-rs over Spark gives developers a broader range of alternatives when working on a "small scale" (1 GB to 100 GB).

1.4 Goals

The accomplishment of the project's purpose (namely, reducing the latency (seconds) and thus increasing the data throughput (rows/second) for reading and writing on Delta Lake tables on HopsFS) is bound to a list of Goals (Gs), here set. These are also related to the set of RQs, outlining a clear structure of the various project milestones.

1. Gs aimed to answer RQ1:

   G1: Understand delta-rs library [29] architecture and dependencies.

   G2: Identify what needs to be implemented to add HDFS support to the delta-rs library [29].

   G3: Implement HDFS support in the delta-rs library [29].

   G4: Verify that HDFS support also works for HopsFS.

2. Gs aimed to answer RQ2:

   G5: Design the experiments to be conducted to evaluate the difference in performance between the current legacy access (Spark-based in writing) to Apache Hudi compared a the delta-rs [29] library-based access to Delta Lake, in HopsFS.

   G6: Perform the designed experiments.

   G7: Visualize the experiments' results, focusing on allowing an effective comparison of performances.

   G8: Analyze and interpret the results in a dedicated thesis report section.

Associated with these Gs several Deliverables (Ds) will be created.

D1: Code implementation adding support to HDFS and HopsFS in the delta-rs library. This D is related to the completion of goals G1–G4. This deliverable also represents the system implementation contribution of the project.

D2: Experiment results on the difference in performance between current legacy access (Spark-based in writing) to Apache Hudi compared a the delta-rs library-based access to Delta Lake, in HopsFS. This D is related to the completion of goals G5–G7.

D3: This thesis document, provides more detail on the implementation, design decisions, expected performance, and analysis of the results. This D is a comprehensive report of all the thesis work, also including the analysis of results defined in G8.

1.5 Ethics and Sustainability

As a systems research project, the focus of this study revolves around software and in particular, developing more efficient data-intensive computing pipelines that find wide application in machine learning and training of neural networks. Software, according to the Green Software Foundation [32], can be "part of the climate problem or part of the climate solution" [33]. We can define Green Software as a software that reduces its impact on the environment by using less physical resources, and less energy and optimizing energy use to use lower-carbon sources [33]. In the context of machine learning and training of neural networks, reducing training time (and so also the read and write latency operation on the dataset) has been proven to positively impact the reduction in carbon emissions [34, 35].

This project, by aiming to reduce the latency (seconds) and thus increase the data throughput (rows/second) for reading and writing on Delta Lake tables on HopsFS, follows the key green software principles reducing Central Processing Unit (CPU) time use compared to the previous pipeline. This leads to a lower carbon footprint, as less energy is being used.

This project contributes to the Sustainable Development Goals (SDGs) 7 (Affordable and Clean Energy) and 9 (Industry Innovation and Infrastructure) [36], more specifically the targets 7.3 (Double the improvement in energy efficiency) and 9.4 (Upgrade all industries and infrastructures for sustainability). This work achieves this by reducing the read and write latency of data on Delta Lake tables, and thus increasing the data throughput. This means that the same amount of data can be read or written in a smaller amount of time, reducing the use of resources (CPU or Graphical Processing Unit (GPU) computing time), thus reducing energy usage. This decrease in energy consumption will lead to a smaller carbon footprint (if the same amount of data is read or written).

Ultimately, this leads to an improvement in energy efficiency and a

Figure 1.1: SDG supported by this thesis

reduction in the carbon footprint of the data-intensive computing pipelines that find wide application in machine learning and training of neural networks.

1.6 Research Methodology

This work starts from a few Industrial Needs (INs), provided by Hopsworks, and a few Project Assumptions (PAs) validated through a literature study. Hopsworks's INs are:

IN1 : the Hopsworks Feature Store using the legacy pipeline (Spark-base in writing) has a high latency (seconds) and low throughput (rows/second) in reading and writing operations when on a "small scale" (1 GB - 100 GB). This highlights the potential for using Spark alternatives in the "small scale" use case.

IN2 : Hopsworks, adapting to their customer needs, supports the Delta Lake table format. Improving the speed of read and write operations on this table format, would improve a typical use case for Hopsworks Feature Store users.

PAs are... These assumptions will be validated in Chapter 2.

PA1 : Python is the most popular programming language and the most used in data science workflows. ML and AI developers prefer Python tools to work. This means that Python libraries with high performance will typically be preferred over alternatives (even more efficient) that are Java Virtual Machine (JVM) or other environments based.

PA2 : Rust libraries have proven to have the chance to improve performance over C/C++ counterparts (Polars over Pandas). A Rust implementation could strongly improve reading and writing operations on the Hopsworks Feature Store.

The project aims at fulfilling the INs with a system implementation approach. First, a HDFS storage support will be written for the delta-rs library to extend the Rust library support to HopsFS [30]. Then, an evaluation structure will be designed and used to compare the performances of the current legacy (Spark-based in writing) system and the new Rust-based pipeline. The two approaches will be tested with datasets of different sizes (between 1 GB and 100 GB). This is critical to identify if the same tool should be used for all scenarios or if they perform differently. The critical metrics that will be used to evaluate the system are read and write operations data throughout (the higher, the better) measured in rows per second. These were chosen as they most affect the computation time of pipelines accessing Delta Lake tables.

1.6.1 Delimitations

The project is conducted in collaboration with Hopsworks AB, and as such the implementation will focus on working with their system using HopsFS. While the consideration drawn from these results cannot be generalized and be true for any system, they can still provide an insight into Apache Spark limitations, and on which tools perform better in different use cases.

1.7 Thesis Structure

Chapter 2 equips readers with the necessary foundational knowledge to navigate the layered data stack that this research operates with. Furthermore, it also introduces the legacy and new system architectures that will be employed during the experiments. Chapter 3 defines the methodology for the two key components of this iterative thesis work, i.e. the system implementation and system evaluation. Chapter 4 details the design choices taken during the system development and describes how to deploy use and the system. Chapter 5 presents the results of the experiments carried out, outlining the differences between the defined pipelines, and the different CPU configurations. The chapter is also complemented with a discussion section that enables the reader to appreciate this thesis findings and implications. Finally, Chapter 6 summarizes the thesis contribution and findings, while also providing a discussion on the limitations and future work.

Chapter 2

Background

This project works on a layered data stack, that handles Big Data, i.e. large volumes of structured and unstructured data types at a high velocity. The data stack handles how data is stored, managed, and retrieved to enable applications built on top of it. As there is no single data architecture that is generally accepted, i.e. different approaches use different architectures [37, 38], thus this project defines a data stack, then focuses on improving its parts, namely the query engine.

The data stack of the project is illustrated in Figure 2.1.

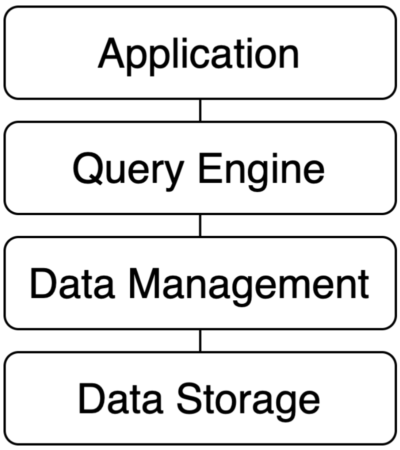

Figure 2.1: Data stack abstraction for this project

The data stack (and this chapter) is divided into four sections:

1. **Data Storage**: handles how the data is stored. The data storage layer might be centralized or distributed, on-premise or on the cloud, storing

data in files, objects, or blocks.

2. **Data Management** : handles how the data is managed. The data management layer might offer ACID properties, data versioning, support open data formats, and support structured and unstructured data.

3. **Query Engine**: handles how data is queried, i.e. accessed, retrieved and written. The query engine might offer caching, highly scalable architectures, and API support for multiple programming languages.

4. **Application**: a system that will take advantage of the data stack. In the case of this project, only the software of the host organization (Hopsworks' Feature Store) will be described.

After explaining the data stack, a section on the legacy and new system architectures complements the Background, showing how the technologies explained are used within the pipelines measured during the experiments.

2.1 Data Storage

This section describes the data storage layer of this project, namely HopsFS. HopsFS is Hopsworks's evolution of HDFS, a distributed file system. HopsFS complexity will be broken down into parts, providing not only a great understanding of the tool but also a comparison with common alternatives, namely cloud object stores.

2.1.1 File storage vs. Object storage vs. Block storage

Data can be stored and organized in physical storages, such as Hard Disk Drives (HDDs) and/or Solid State Drives (SSDs), in three major ways: (1) Files, (2) Objects, and (3) Blocks. For each one, the technique is briefly described and then a comparative table is shown (Table 2.1). This subsection is a rework according to the author's understanding of three articles coming from major cloud providers (Amazon, Google, and IBM) [39, 40, 41].

File storage

File storage is a hierarchical data storage technique that stores data into files. A file is a collection of data characterized by a file extension (e.g. ".txt", ".png", ".csv", ".parquet") that indicates how the data contained is organized. Every file is contained within a directory, that can contain other files or other directories (called "subdirectories"). In many file storage systems directories are called "folders".

This type of structure, very common in modern Personal Computers (PCs), simplifies locating and retrieving a single file and its flexibility allows it to store any type of data. However, its hierarchical structure requires that to access a file, its exact location should be known. This restriction decreases the scaling possibilities of the system, where a large amount of data needs to be retrieved at the same time.

Overall, this solution is still vastly popular in user-facing storage applications (e.g. Dropbox, Google Drive, One Drive) and PCs thanks to its intuitive structure and ease of use. On the other hand, other options are preferred for managing large quantities of data, due to its lack in scalability.

Advantages

- Ideal for small-scale operations (low latency, efficient folderization).
- User familiarity and ease of management.
- File-level access permissions and locking capabilities.

Disadvantages

- Difficult to scale due to deep folderization.
- Inefficient in storing unstructured data.
- Limitations in scalability due to reaching device or network capacity.

Object storage

Object storage is a flat data storage technique that stores data into objects. An object is an isolated container associated with metadata, i.e. a set of attributes that describe the data e.g. a unique identifier, object name, size, and creation date. Metadata is used to retrieve the data more easily, allowing for queries that retrieve large quantities of data simultaneously, e.g. all data that was created on a specific date.

The flat structure of Object storage, where all objects are in the same container called a bucket, is ideal for managing large quantities of unstructured data (e.g. videos, images). This structure is also easier to scale as it can be replicated easily across multiple regions allowing faster access in different areas of the world, and fault tolerance to hardware failure.

On the other hand, objects cannot be altered once created, and in case of a change must recreated. Also, object stores are not ideal for transactional operations as objects cannot have a locking mechanism. Lastly, object stores have slower writing performance compared to file or block storage solutions.

Overall, this solution is widely when high scalability is required (e.g. social networks, and video streaming apps) thanks to its flat structure and use of metadata. On the other hand, other options are preferred when transactional operations are required, or high performance on a small number of files that change frequently is necessary.

Advantages

- Potential unlimited scalability.
- Effective use of metadata enabling advanced queries.
- Cost-efficient storage for all types of data (also unstructured).

Disadvantages

- Absence of file locking mechanisms.
- Low performance (increased latency and processing overhead).
- Lacks data update capabilities (only recreation).

Block storage

Block storage is a data storage technique that divides data into blocks of fixed size that can be read or written individually. Each block is associated with a unique identifier and it is then stored on a physical server (note that a block can be stored in different Operating Systems (OSes)). When the user requests the data saved, the block storage retrieves the data from the associated blocks and then re-assembles the data of the blocks into a single unit. The block storage also manages the physical location of the block, saving a block where is more efficient.

Block storage is very effective for systems needing fast access and low latency. This architecture is compatible with frequent changes, unlike object storage.

On the other hand, block storage achieves its speed by operating at a low level on physical systems, so the cost of the architecture is strictly bound to the storage and servers used, not allowing the architecture to scale according to its demands.

Advantages

- High performance (low latency).
- Reliable self-contained storage units.
- Data stored can be modified easily.

Disadvantages

- Lack of metadata brings limitations in data searchability.
- High cost to scale the infrastructure.

Table 2.1: Data storage features comparison

Characteristics	File Storage	Object Storage	Block Storage
Performance	High	Low	High
Scalability	Low	High	Low
Cost	High	Low	High

2.1.2 Hadoop Distributed File System

Hadoop Distributed File System (HDFS) is a Distributed File System (DFS), i.e. a file system (synonym of file storage) that uses distributed storage resources while providing a single namespace as a traditional file system. HDFS has significant differences compared with other DFSes. HDFS is highly fault-tolerant, i.e. it is resistant to hardware failures of part of its infrastructure and can be deployed on commodity hardware. HDFS also provides high throughput access to application data and it is designed to be highly compatible with applications with large datasets (more than 100 GB) [42].

HDFS architecture consists of a single primary node called Namenode and multiple secondary nodes called Datanodes. The Namenode manages the filesystem namespace and regulates access to files by clients. On the other hand, Datanodes manage the storage attached to the nodes they run on and they are responsible for performing replication requests when prompted by the Namenode. HDFS exposes to users a file system namespace where data can be stored in files. Internally, a file is divided into one or multiple blocks and these blocks are stored in a set of Datanodes. The blocks are also replicated upon the first write operation, up to a certain number of times (by default three times, with at least one copy on a different physical infrastructure). The Namenode keeps track of the data location, matching it with the filesystem namespace. It is also responsible for managing Datanode reachability (through periodical state messages sent by Datanodes), and providing clients with the locations of the Datanodes containing the blocks that compose the requested file. If a new write request is received, it is still the Namenode that needs to provide the locations of available storage for the file blocks.

In Figure 2.2 a simplified visual representation of the Namenode and Datanodes basic operations in HDFS is present.

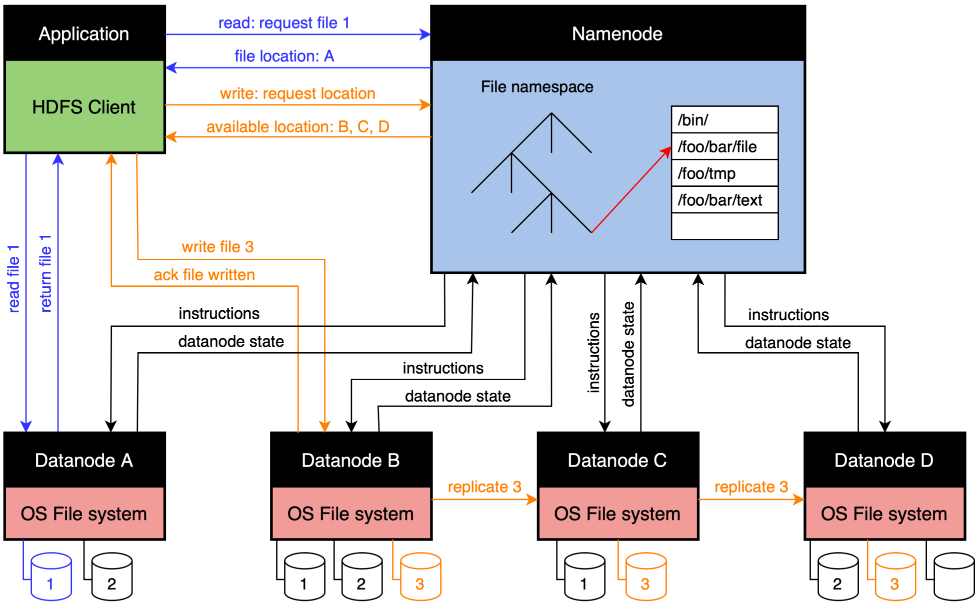

Figure 2.2: HDFS architecture during read and write operations

2.1.3 HopsFS as an HDFS evolution

HopsFS is HDFS improved release first presented at the 15th USENIX conference in 2017 [30]. HopsFS stores the metadata in an external NewSQL database with high throughput and low operation latency called RonDB [43]. This enables this storage solution to avoid having all metadata in a single Namenode, but allowing for Namenode replication, having the same metadata on RonDB. This solution proved to have from sixteen up to thirty-seven times the performance on HDFS thanks to its ability to scale according to metadata being stored, and also on similar setups where the same resources are being utilized.

HopsFS is one of the core technologies on which the Hopsworks Feature Store is built. While the system has not seen more widespread use, it is extremely impactful in its current applications, contributing to making Hopsworks "outperform existing cloud feature stores for training and online inference query workloads" [44].

2.1.4 HDFS alternatives: Cloud object stores

Thanks to its high scalability and low cost (Section 2.1.1) object storage has been widely adopted to store large quantities of unstructured data. Starting with Amazon Web Services (AWS) in 2006 with its S3 service, a lot of other vendors started offering cloud object storage services. The main advantages of using cloud resources are related to their elasticity that combined with object storage capability enable users to use as much storage as they need and be billed for what was used.

HDFS was first released in 2006, and so it evolved in parallel with cloud objects storage solutions. While HDFS was widely adopted for on-premise solutions, more and more businesses migrated their operations to cloud services. Nowadays, cloud object storage like AWS's S3, Google's Google Cloud Storage (GCS) and Microsoft's Azure are widely used and libraries, e.g. delta-rs, prioritize support for these platforms as they are widely adopted. HDFS and its evolutions, e.g. HopsFS, still see use, but it fits more specific use cases, as the convenience of a cloud service is highly valued by the market.

2.2 Data Management

This section describes the data management layer for this project, i.e. how data is managed, versioned, etc. The main DBMS technology used in this is Delta Lake, and to understand it Section 2.2.1 revises the problems that technologies aimed to solve, and the limitations of these systems. The chapter is complemented with Section 2.2.2 which explains how to access Delta Lake and introduces delta-rs.

2.2.1 Brief history of Data Base Management Systems

In recent years the rise of Big Data, large volumes of various structured and unstructured data types at a high velocity, has shown an incredible potential but it has also posed several challenges [45]. These mostly impact the software architecture that needs to deal with these issues, which led to an evolution of these technologies [46]. Delta Lake [7], is one of the most recent iterations of this evolution process, but to understand the tool, it is necessary to understand the challenges, starting from the beginning of the data management evolution.

Before Big Data, companies already wanted to gain insights from their data sources using an automated workflow. Here is where ETL and relational databases first came into use. An ETL pipeline as the name suggests:

1. Extracts data from APIs or other company's data sources.
2. Transforms data by removing errors or absent fields, standardizes the format to match the database, and validates the data to verify its correctness.
3. Loads it into a relational database (e.g. MySQL).

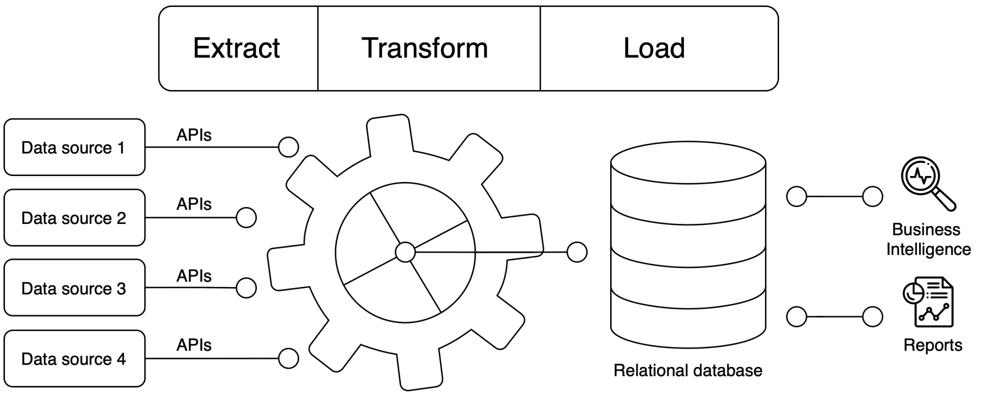

Figure 2.3: Simple ETL system with a relational database

This type of workflow, represented in Figure 2.3, enabled companies to obtain BI insights and data reports on the company's data. The main limitation of this system sat in its limited capability of creating reports or BI insights based on data sitting on multiple tables. These types of requests are called analytical queries, and while they might run less often than simpler queries are still crucial for making data-driven decisions (e.g. determining the region that sold more product units in the last year).

When the need to compute analytical queries rose, more complex DBMS substituted the simple relational databases, optimizing for running business-centric complex analytical queries. These systems are called OLAP, and its prime example example is the data warehouse.

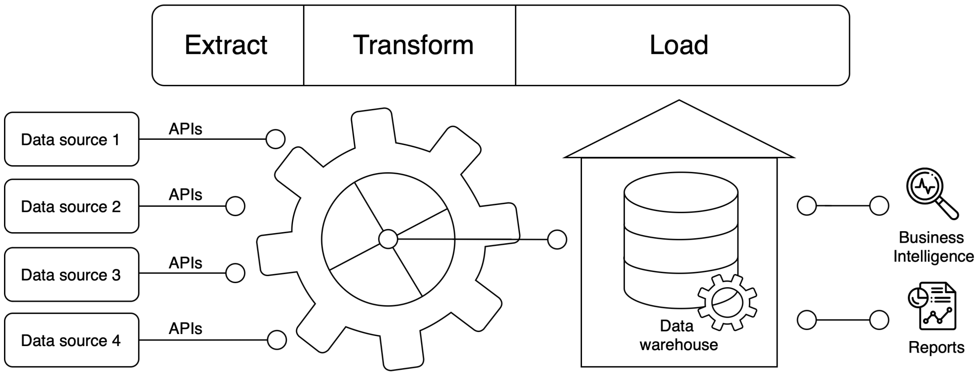

Figure 2.4: ETL system with a data warehouse

A data warehouse workflow, visualized in Figure 2.4, enables larger quantities of data to be computed and analyzed. Data warehouses enable BI and Reports that consider all data sources and can join multiple tables efficiently. This type of DBMS still keeps a relational database key features, such as ACID transactions and data versioning.

Over time, the rapidly growing amount of unstructured data (also called Big Data, e.g. images, and videos) created new needs within companies, that wanted to take advantage of this new data. Data warehouses were unfit to solve this problem as they only supported structured data. Furthermore, storing large quantities of data in data warehouses is expensive and does not support any type of AI/ML workflow.

These issues were tackled by a new paradigm called Data Lake (Figure 2.5). Data Lakes are based on a low-cost object storage system (Section 2.1.1) that consists of a flat structure where all data is loaded after extraction. In data lakes the architecture structure changes as data is first loaded into the data lake and only after transformed. This paradigm is called ELT. Transformations are customizable for specific applications, e.g. BI and reports using a Data warehouse, an AI/ML analysis.

This architecture reduces storage costs but also increases the system

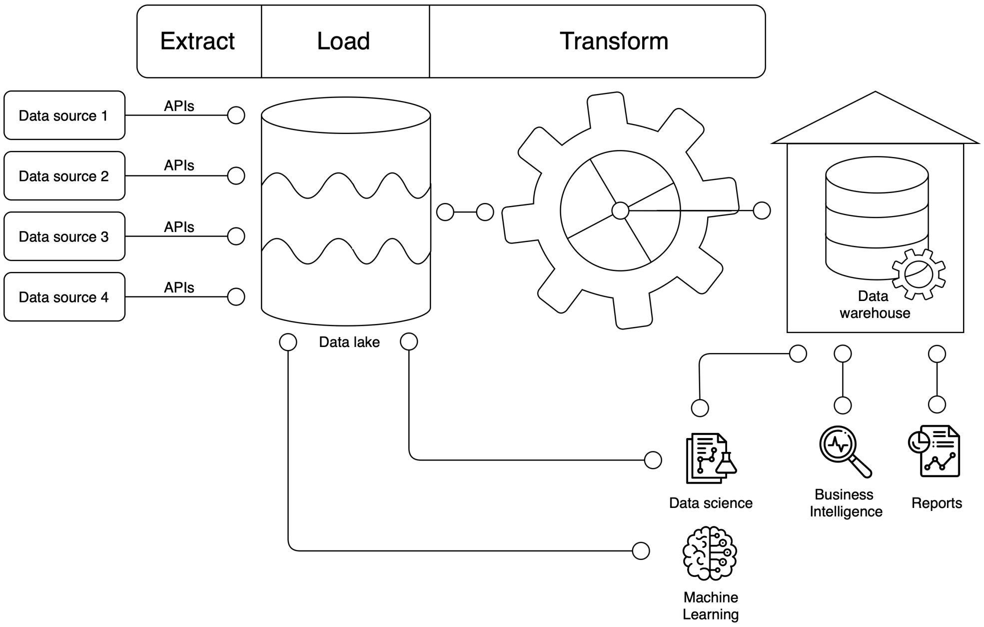

Figure 2.5: ELT system with a data lake

complexity. Higher complexity is typically related to higher costs, as system maintenance is more costly and can lead to a larger number of issues. Additionally, since data lake cannot be queried directly with BI queries or requesting business reports, this leads to the need to still maintain a data warehouse (as in Figure 2.5). This ultimately leads to higher costs to maintain the multiple storages for the same data. The system also suffers from timeliness due to this lengthy pipeline, as the data needs to go through many steps before it is available in the data warehouse.

These issues outlined that data lakes were not a drop-in replacement for data warehouses, as they served a different purpose and suffered from different issues. This generated the need to have a system that could have data warehouses ACID and data management capabilities while being able to support unstructured data. The solution was a new architecture, the data lakehouse.

Figure 2.6 shows a Delta Lake system [7]. The data lakehouse term was first introduced by Databricks in 2021 with their paper presented at Conference on Innovative Data Systems Research (CIDR) [2], while data lakehouse solutions already existed on the market, namely Apache Hudi [5], Apache Iceberg [6] and Delta Lake [7]. Data lakehouses combine the benefits of data warehouses and data lakes while simplifying the complexity of storing

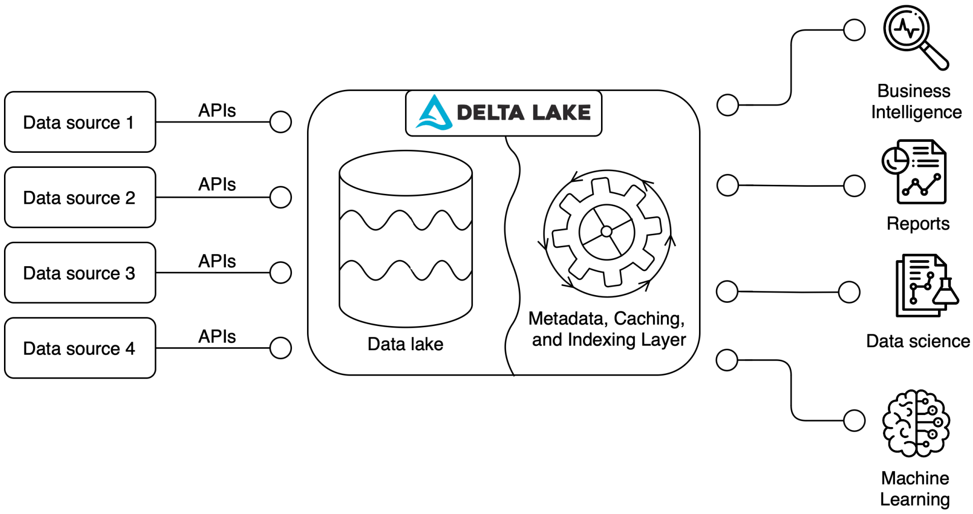

Figure 2.6: Delta lake system

and accessing data in enterprise data architectures. This new architecture is based on an open-data format called parquet [47], which is a column-based data file format. Data is saved in a data lake in the form of parquet files, then a management layer enables managing transactions, versioning, indexing, and other data management features. This enables data lakehouses to offer ACID properties, similarly to data warehouses.

Delta Lake is the technology that will be most used in this project. A Delta Lake Table (i.e. an instance of Delta Lake) works operating on a data lake containing parquet files and a transaction log. The transaction log records each operation, enabling versioning and recovery of previous versions (also called time travel). The data lake can be partitioned, e.g. using a date field, and this would make the Delta Lake table look as Figure 2.7.

2.2.2 Accessing Delta Lake

The Delta Lake project is strictly related to Apache Spark, as Databricks (the company that developed Delta Lake) was built by the Spark developers [24] to offer big data management services around Spark. As of version 3.0 of Delta Lake, Delta Kernel was announced [48], a Java library providing low-level access to Delta Lake, without needing to write the Delta Lake logic. While this move tried to standardize all accesses to Delta Lake under the Spark/Java environment, new ways to access the library had already been written since

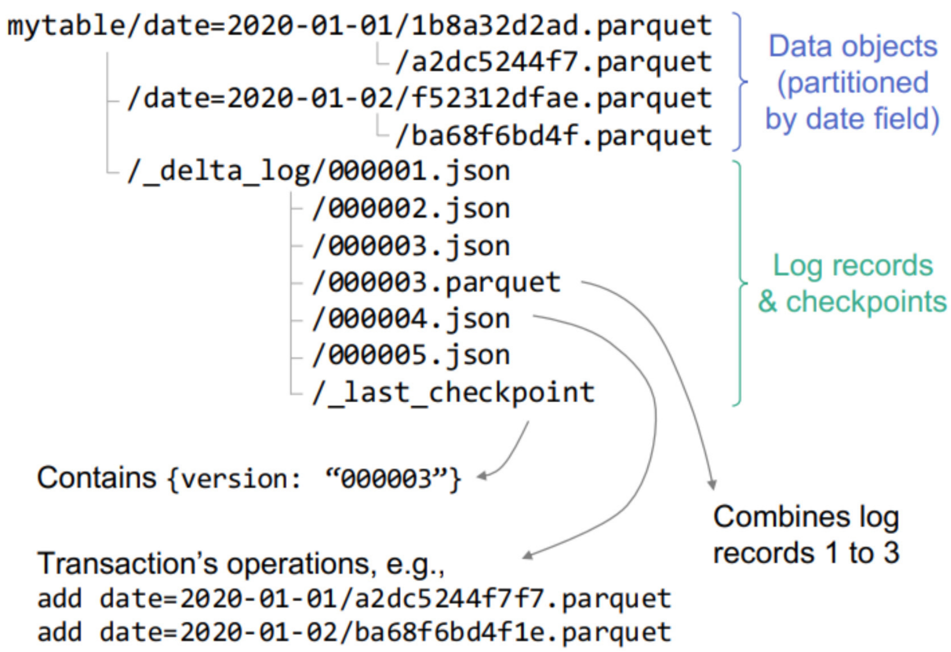

Figure 2.7: Delta lake table partitioned according to date field

Delta Lake is an open-source project.

As early as April 2020, the Rust community started implementing a new interface to access Delta Lake written in Rust without JVM dependencies. This library expanded even more the Delta Lake ecosystem, breaking the dependency between Delta Lake and Spark to perform operations on the data lakehouse. This worked particularly well considering that the data science community makes heavy use of Python and typically would avoid having JVM dependencies. Being delta-rs written in Rust it makes the library highly compatible with Python since it can be easily wrapped and deployed as a Python library. This is perhaps the reason behind the fact that delta-rs can be simply installed in Python with the name "deltalake".

2.3 Query engine

This section described the technologies used to query, cache, and process data in this project. Query engine refers mainly to Apache Spark and DuckDB, but also the other technologies presented in this chapter operate at the same abstraction level while having different functions.

2.3.1 Apache Spark

Apache Spark (from now on simply Spark) is an open-source distributed computing framework designed to handle large-scale data-intensive applications [8]. Spark builds from the roots of MapReduce and its variants. MapReduce is a distributed programming model first designed by Google that enables the management of large datasets [49]. The paradigm was later implemented as an open-source project by Yahoo! engineers under the name of Hadoop MapReduce [42]. Spark improved this approach by making use of Resilient Distributed Datasets (RDDs) [50]. RDDs are a distributed memory abstraction that enables a lazy in-memory computation that is tracked through the use of lineage graphs, ultimately increasing fault tolerance [50]. The difference between Hadoop MapReduce and Spark is represented in Figure 2.8.

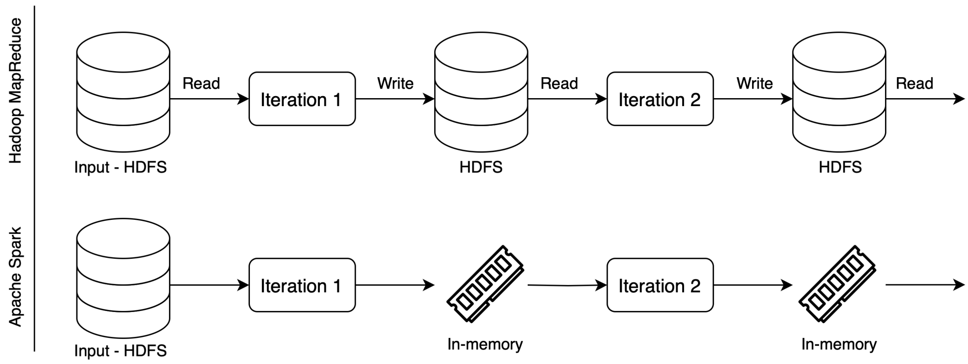

Figure 2.8: Hadoop MapReduce and Apache Spark execution differences

2.3.2 Apache Kafka

Apache Kafka (from now on, Kafka) is an open-source distributed data streaming platform designed for high-throughput, and scalable data processing [51]. Kafka is most typically used for real-time streaming applications thanks to low latency.

Figure 2.9 shows the components and messages exchanged in a Kafka cluster. The key components of Kafka are:

1. **Producer**: an application that produces and labels data with specific topics (shapes in the figure).

2. **Zookeeper**: responsible for keeping track of brokers and the topic's current offset.

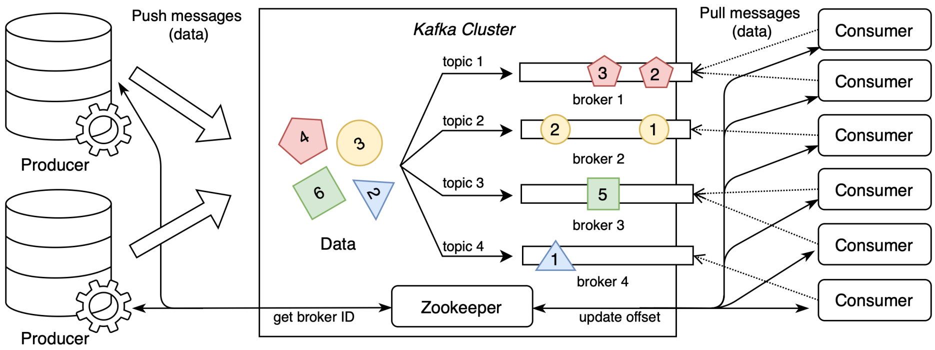

Figure 2.9: Kafka cluster

3. **Broker**: a computational node (or server) that handles data on a specific topic. It is responsible for receiving messages from producers. Once messages are received the broker forwards them on request by the consumers. This enables the asynchronous protocol.

4. **Consumer**: an application that is interested in topic-specific data. To access data it is subscribed to a topic receiving new messages when published. To a topic, more consumers can be subscribed.

The protocol allows applications acting as producers and consumers to avoid having a synchronous protocol. This enables producer's high throughput since they can send messages without waiting for consumers to process them. This also allows consumers to be flexible on workload size.

Given its distributed nature, Kafka allows several producers, consumers, and brokers to exist at the same time. This enables the system to be tuned according to the needs of a specific application.

2.3.3 DuckDB

DuckDB [10] is an open-source, embedded, OLAP DBMS. DuckDB was designed to process small quantities of data (1 - 100 GBs) within the same process or application that runs it, instead of a different process/application. These features create an efficient OLAP database that can be used for data analysis, and data processing on a small scale, without the complexity of a more complex DBMS, e.g. Teradata [52].

The light structure that characterizes DuckDB is what enables this system to be extremely responsive with low latency. The limitations of the system regard data size, as DuckDB makes use of in-memory processing, it is not able to handle big workloads (1TB or more), which require multiple disk loads.

2.3.4 Arrow Flight

Arrow Flight is a high-performance framework for data transferring over a network, most typically Arrow tables [53]. This protocol enables the transfer of large quantities of data stored in a format, e.g. Arrow tables, without having to serialize or deserialize it for transfer. This speeds up the data transfer by a large margin making Arrow Flight extremely efficient. Arrow Flight is designed to be cross-platform, having support for multiple programming languages (C++, Python, Java). The protocol also supports parallelism, speeding up transfers by using multiple nodes on parallel systems. Arrow Flight protocol is built on top of gRPC [54], enabling standardization and an easier development of connectors.

2.4 Application - Hopsworks Feature Store

This section describes the application layer which takes advantage of the data stack described. The software described is the Hopsworks Feature Store, which this project contributes to.

2.4.1 MLOps fundamental concepts

Machine Learning Operations (MLOps) are a set of practices to automate and simplify ML workflows from the data collection, to the model deployment. MLOps considers the problem of developing and deploying a ML system from a code point of view and data point of view. While for a typical software application, only code needs versioning, for a ML application also data needs versioning, as training on different data versions might affect performance. The need for data versioning saw Feature Store emerge as a solution for the problem [55]. Feature stores play a central role in the MLOps process, serving as fast-access storage during the process.

Figure 2.10 shows a simple MLOps architecture making use of the Hopsworks' AI data platform. After data is gathered from various sources, a feature pipeline processes the data performing model-independent transformations and saving the resulting features in the Feature store. The training pipeline then runs model-dependent transformation (based on the specific model that is going to be trained), on the same features retrieved from the Feature Store and saves the output, i.e. the model in a Model registry. Then a last process is responsible for performing inference, and it is typically embedded in deployed applications. For this process, it will be enough to take the new features gathered on the platform and perform an inference on the specific model saved in the Model registry.

This type of architecture allows an asynchronous decoupled pipeline that enables the system to be maintainable, scalable, and extremely effective for production scenarios.

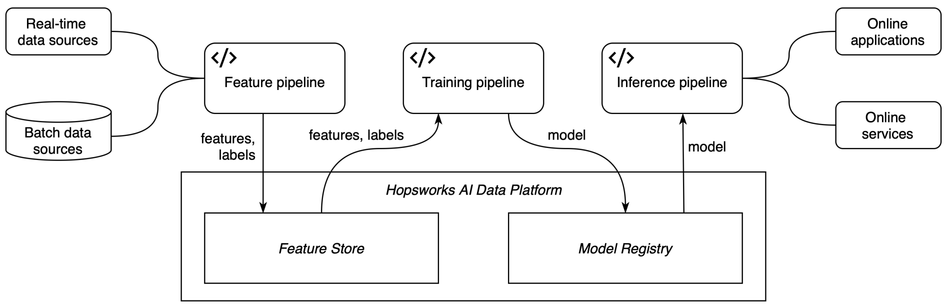

Figure 2.10: MLOps pipeline using a Feature Store and a Model Registry

2.4.2 Hopsworks Feature Store

As first introduced in the previous section, the Feature Store is a key data layer in an MLOps workflow. Feature Store enables feature reusability, a centralized and easier collaboration on model training and deployment. The Hopsworks Feature Store organizes features in feature groups, i.e. a mutable collection of features. Feature groups can be queried via the Hopsworks API allowing developers to perform Create Read Update Delete (CRUD) operations.

The Hopsworks Feature Store in addition to supporting batch data sources also supports real-time data streaming. To be able to support both systems (or hybrids), the Hopsworks Feature Store saves features in two storages: the Offline Feature Store and the Online Feature Store. The Offline feature store is a column-based storage suited for batch data, that is updated with a low frequency (every few hours at maximum frequency). The Online feature store on the other hand is a row-based, key-value, data storage based on RonDB [43]. These characteristics enable low latency and real-time (in seconds) data processing. To keep this dual system consistent the Hopsworks Feature Store has a unique point of entry for data which is Kafka, that guarantees at least one message delivery to both storages. This enables the system to support both workflows while keeping consistent data storage.

2.5 System architectures

This section describes the architectures of the legacy and new systems that will be run and measured in the experimental part of this thesis. The section is divided into four sections and each schema presented shows the protocol step by step.

2.5.1 Legacy system - Writing

Figure 2.11 (Figure A.1 in larger size) shows the legacy Hopsworks Feature Store write process from the client onto the Offline Feature Store. The process is mainly split into two synchronous parts: upload and materialization. In the upload step, the Pandas data frame given as input is converted into rows and sent one row at a time to Kafka. Then, when the upload is finished the client is notified. Asynchronously, a Spark job has been running in the cluster since the Hopsworks cluster was started, which is the Hudi Delta Streamer. This job periodically retrieves messages from Kafka, and then once it retrieves a full table it writes it in a column-oriented format to Apache Hudi, which sits on top of a HopsFS system. Once the materialization is completed the Python client is also notified of completion.

As in the pipeline, the upload and the materialization are two different parts of the process that do not act synchronously. During the experimental part of the thesis, to be able to measure the latency of the whole process without having to account for the Hudi Delta Streamer data retrieval period, the materialize function was called, which allows the system to perform the materialization on call instead of waiting for the period. This enabled the experiments to retrieve accurate data on the total latency of the process.

2.5.2 Legacy system - Reading

Figure 2.12 (Figure A.2 in larger size) shows the legacy Hopsworks Feature Store read process from the client onto the Offline Feature Store. The process, differently from the writing process, is not Spark-based and it is using a Spark alternative: a combination of an Arrow Flight server and a DuckDB instance.

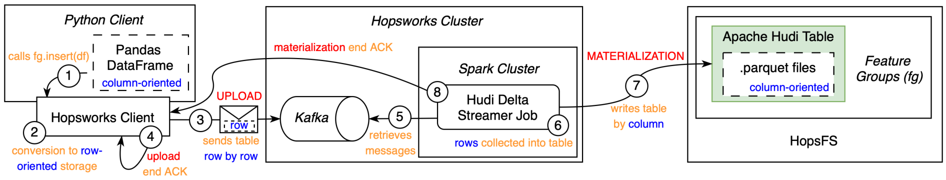

Figure 2.11: Legacy system writing process

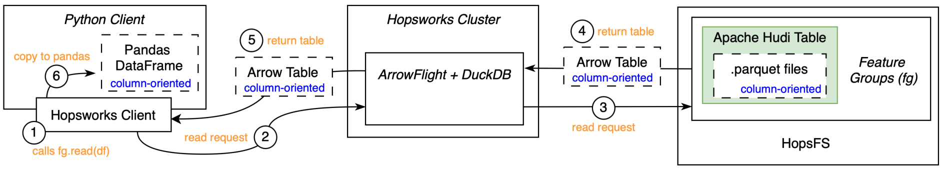

Figure 2.12: Legacy system reading process

This avoids the serialization and deserialization into row-based tables for sending the data, keeping the unified standard Arrow Table, which is a column-oriented format.

2.5.3 New system - Writing

Figure 2.13 shows how the delta-rs library writes on a Delta Lake table instanced on top of HopsFS. The delta-rs library streamlines the process, without having to pass from a server instance (Spark).

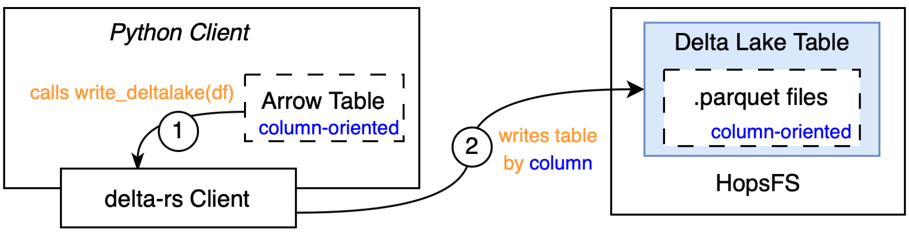

Figure 2.13: delta-rs library writing process

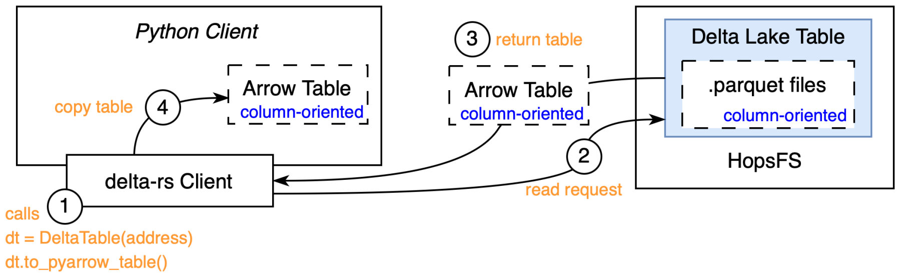

Figure 2.14: delta-rs library reading process

2.5.4 New system - Reading

Figure 2.14 shows how the delta-rs library reads on a Delta Lake table instanced on top of HopsFS. The delta-rs library streamlines the process, without having to pass from a server instance (Arrow Flight).

Chapter 3

Method

This chapter following the two RQs defined in Section 1.2.1 defines two methodologies that will be applied sequentially in this project. Section 3.1 defines the system implementation process which outputs the D1, i.e. the code implementation, that will enable the system evaluation defined in Section 3.2 which will output D2, i.e. the results of the experiment which analysis will be delivered in D3.

3.1 System implementation – RQ1

The core of this system research thesis resides in the system implementation section which answers RQ1.

This section explains the method and the principles that will be used to carry out the software development process of adding support for HDFS and HopsFS to the delta-rs library. This section is divided into four sub-sections: Research paradigm, explaining the research framework that will be used to implement the system, Development process, explaining the activities that will be carried out to implement the system, Requirements, defining the functional and non-functional requirements of the system and Development environment, detailing the tools and resources that will be used during the development process.

3.1.1 Research Paradigm

The research paradigm of the system implementation section of this thesis is positivist, believing that reality is certain and it can be measured and understood. This thought declined in the context of the development process, which means that the new system requirements can be defined and implementation errors can be outlined, understood and if possible fixed. This approach leads to a strict definition of the development process depending on functional requirements that must be fulfilled.

3.1.2 Development process

The development process will follow an iterative and incremental development approach described in Figure 3.1. This methodology will be applied as it allows flexibility while creating incrementally a working system [56]. This project, due to the need to work on HopsFS, will require numerous interactions with HopsFS maintainers (i.e. the industrial supervisor). This creates the need for a feedback loop, which will allow the system to fit all the requirements according to all stakeholder's expectations.

As it can be noted from Figure 3.1. Each step of this process is related to one of the goals (G1–G4) associated with RQ1 in Section 1.4. The activities and the relationship between each activity and the associated goal(s) are here explained:

1. **Identify problems collaboratively**: this activity solves partially G1–G2, as it is an initial system analysis, performed together with the industrial supervisor, who is knowledgeable on Hopsworks' infrastructure (in particular HopsFS). This task fixes the initial requirements of the project and investigates what needs to be implemented at a high-level abstraction.

2. **Analyse system**: this activity solves partially G1–G2 each time is re-iterated, as it performs low-level code-based analysis of how the system works and what needs to be implemented to support HDFS in delta-rs. This activity also starts an iterative loop that will end once the system fulfills the requirements described in Section 3.1.3.

3. **Design software**: this activity solves partially G3, as the first part of the software development. In this activity, the system analyzed before is considered to design a solution.

4. **Code software**: this activity solves partially G3, as the second part of the software development. In this activity, the solution designed is coded.

5. **Test system**: this activity solves partially G4, as the first part of the tests performed to verify the solution validity, via unit tests. Failed unit tests

will trigger a new development loop iteration, where this failure will be considered as the first starting point in the system analysis.

6. **Verify system integration**: this activity solves partially G4, as the second part of the tests performed to verify the solution validity, via integration tests. Failed integration tests will trigger a new development loop iteration, where this failure will be considered as the first starting point in the system analysis. On the other hand, if the integration test succeeds, the loop will be restarted if the system does not yet fit a requirement, or finished if the system fulfills all requirements described in Section 3.1.3.

This process will produce a final deliverable (D1), which is a Python wheel of the delta-rs library containing the support for HDFS and HopsFS.

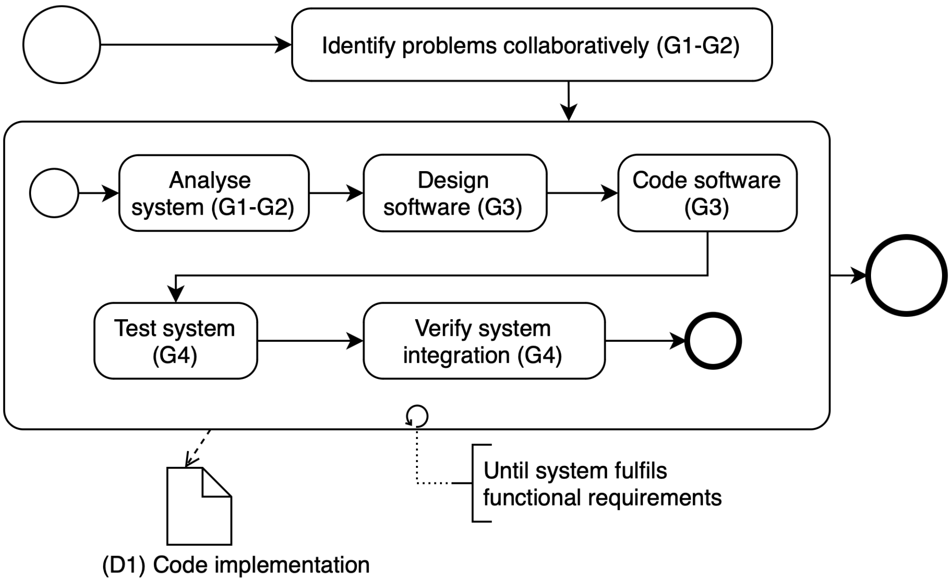

Figure 3.1: BPMN diagram of the System implementation process

3.1.3 Requirements

In the first steps of the system analysis, a series of requirements are defined in agreement with the industrial partner Hopsworks AB, to favor the creation of a solution that could be later used within the company, in a production

environment. These are divided into two categories: functional and non-functional requirements.
The **functional requirements** are:

1. **Write Delta Tables**: the solution should allow to write Delta Lake tables on HopsFS via the delta-rs library.

2. **Read Delta Tables**: the solution should allow to read Delta Lake tables on HopsFS via the delta-rs library.

3. **Communicate via TLS**: the solution should interact with HopsFS via Trasport Layer Security (TLS) protocol version 1.2.

The **non-functional requirements** are:

1. **Consistent**: the solution should be consistent with the current open-source codebase used when appropriate.

2. **Maintainable**: the solution should minimize the need for maintenance and support of the codebase in the future, minimizing changes to open-source code. When appropriate, the changes the solution introduces should be compatible with a future upstream merge to the open-source project modified.

3. **Scalable**: the solution should be able to handle larger quantities of data (up to 100 GB) read or written on Delta Tables.

3.1.4 Development environment

The system implementation will be developed by making use of several technologies, here categorized:

- **Computing resources**: the system implementation will be developed in a remote environment accessed via Secure Shell protocol (SSH) from a computer terminal. This remote Virtual Machine (VM) is selected as mounting HopsFS on a local machine is non-trivial and developing locally could result in inconsistencies when the solution is reproduced in a virtual environment.

- **Writing code**: the Vim [57] text editor is development tool of choice in combination with Conquer of Completion (CoC) [58] providing language-aware autocompletion and rust-analyzer [59] access for on-code compiler errors.

- **Libraries and dependencies**: for simpler development and test reproducibility, the environment will be set in a Docker container [60].
- **Code versioning and shared development**: GitHub [61] will be used for versioning, collaborating with open-source projects (e.g. delta-rs), and sharing the developed solution.

3.2 System evaluation – RQ2

The system evaluation complements the system implementation by measuring the performance of the developed solution, answering RQ2. This evaluation process will be carried out following a sequential approach.

This section details the research paradigm, the method, and the principles that will be used to carry out the system evaluation process, measuring the performance (latency, measured in seconds, and throughput, measured in rows/second) of reading and writing on Delta Lake or Apache Hudi while on HopsFS of the current legacy pipeline (Spark-based in writing) and Rust-based pipelines.

3.2.1 Research Paradigm

The research paradigm for the system evaluation section of this thesis has a more hybrid approach between positivism and post-positivism. While still performing a confirmatory research approach, based on defined objectives, it considers the limitations and biases of this approach, not seeking to generalize the results obtained to other cases. This approach is motivated by the limitations and biases given by the industrial context of this thesis while performing a confirmatory analysis based on a newly implemented system.

3.2.2 Evaluation Process

The evaluation process will follow a sequential approach described in Figure 3.2. Each step of this process is related to one of the goals (G5-G8) associated with the RQ2 in Section 1.4. The relationship between each activity and the associated goal(s) is here explained:

1. **Design experiments**: this activity maps perfectly to G5, designing the experiments that will be conducted to evaluate the performance

difference in performance between the current legacy access (Spark-based in writing) to Apache Hudi compared a the delta-rs [29] library-based access to Delta Lake, in HopsFS.

2. **Perform experiments**: this activity maps perfectly to G6, using the code implementation (D1) to conduct the designed experiments on the analyzed systems. Here data is collected as latency expressed in seconds.

3. **Transform data according to metrics**: this activity is requisite to fulfill G7, as throughput is computed from latency and not measured. The relationship that relates throughput (rows/second), latency(seconds), and size of table (rows) is the following:

$$throughput\ (rows/second) = \frac{number\ of\ rows\ (rows)}{latency\ (seconds)}$$

4. **Visualize results**: this activity maps perfectly to G7, visualizing the experiments' result according to two metrics, i.e. latency measured in seconds and throughput measured in rows/second. This activity also generates a deliverable (D2) composed of the experiment results complete with tables and histograms, i.e. Chapter 5.

5. **Analyze results**: this activity maps perfectly to G8, analyzing and interpreting the results delivered in D2. This contributes to D3, generating the analysis of results, i.e. Chapter 5.

3.2.3 Industrial use case

For the system evaluation to be performed several choices must be made to select: which data is going to be used, which environment will run the experiments, and which metrics will be used to evaluate the system. While the other sections of this chapter detail which decisions were taken, this section aims to outline a typical use case of the system implementation that can motivate the choices between how the system is going to be evaluated.

During the research work in Hopsworks AB, the author had the chance to talk to various employees and form an idea of what a typical client use case for the Hopsworks Feature Store looks like. While these parameters are qualitative, they depict a specific use case around which this thesis work was built. Here below are these use case characteristics:

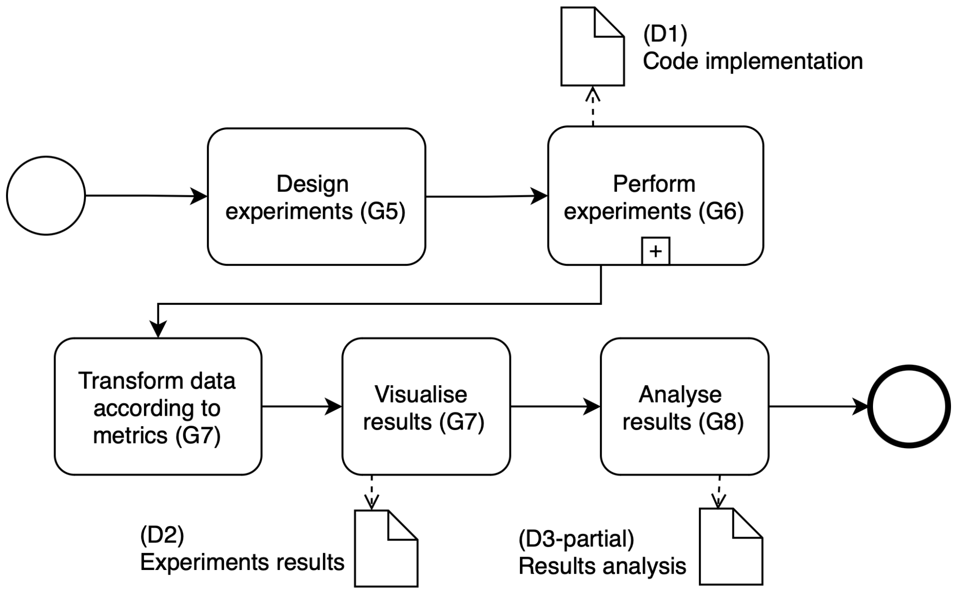

Figure 3.2: BPMN diagram of the System evaluation process

- **Usage of rows over storage size**: Contrary to the introduction chapter where the size of workload is mostly referred to by storage size (bytes), in the experimental part of the thesis only the number of rows will be used to refer to size. This is motivated by the need to find a reliable unit that measures a table size. Storage (bytes) is not strictly linked to a table structure (a table could have a lot of columns and a small number of rows and occupy the same memory as a table with a lot of rows but few columns), and it varies across platforms (different DBMS might save information with different overheads) and storage structures (row-oriented format vs. column-oriented format, i.e. csv vs. parquet). Thus storage size (bytes) alone is unreliable and cannot be used to measure data pipeline performance. In this thesis, number of rows was kept as the main unit to measure data size, but it was also supplemented by specifying the number of columns and the data types contained in each row in Section 3.2.4.

- **Selected table size**: within Hopsworks the author had the chance to see that most of the company's clients' workloads were limited (from 1M to 100M rows), while few clients had massive workloads (more than 1 BN rows). Given this outlook, this project opted to work on improving performance for the smaller workloads setting the table sizes between 100K and 60M rows. This is further detailed in Section 3.2.4.

- **Type of data**: the Hopsworks Feature Store works only with structured data (e.g. numbers, strings), and not with unstructured data (e.g. images, videos, audio) so also the selected dataset in Section 3.2.4 will reflect this scenario.

- **Client configuration**: the performance of the implemented system also depends on the computational and storage resources available on the client configuration running the system. To reflect a typical use case scenario, four CPU configurations were selected (1 core, 2 cores, 4 cores, and 8 cores), while Random Access Memory (RAM) memory will be adapted to the need of the system (as it is unknown before running the experiments). Additionally, to avoid storage I/O bottlenecks and reflect a modern system, a system equipped with a SSD storage will be used. This is further detailed in Section 3.2.6.

3.2.4 Dataset

The data that will be used to perform the read and write operations comes from TPC-H benchmark suite[62]. TPC-H is a decision support benchmark by Transaction Processing Performance Council (TPC). It consists of a series of business-oriented ad-hoc queries designed to be industry-relevant [63]. The data coming from this benchmark suite was used as it provides a recognized standard for data storage systems [64], and it has already been used in similar related work [10, 65]. Any part of the data can be generated using the TPC-H data generation tool [66].

The TPC-H benchmark contains eight tables, of these two, SUPPLIER and LINEITEM, were selected for the following reasons. The two tables are respectively the smallest (10000 rows) and largest (6000000 rows) tables whose size (number of rows) depends on the Scale Factor (SF). The SF can be varied to obtain tables of different sizes (number of rows), allowing a progressive change in the table size (number of rows).

The SUPPLIER table has seven columns, while the LINEITEM table has sixteen. This influences the average size of memory each row occupies. Here below for each table its columns, with their specific type, are listed.

- SUPPLIER
  - S_SUPPKEY : identifier

    - S_NAME : fixed text, size 25
    - S_ADDRESS : variable text, size 40
    - S_NATIONKEY : identifier
    - S_PHONE : fixed text, size 15
    - S_ACCTBAL : decimal
    - S_COMMENT : variable text, size 101

- LINEITEM
    - L_ORDERKEY : identifier
    - L_PARTKEY : identifier
    - L_SUPPKEY : identifier
    - L_LINENUMBER : integer
    - L_QUANTITY : decimal
    - L_EXTENDEDPRICE : decimal
    - L_DISCOUNT : decimal
    - L_TAX : decimal
    - L_RETURNFLAG : fixed text, size 1
    - L_LINESTATUS : fixed text, size 1
    - L_SHIPDATE : date
    - L_COMMITDATE : date
    - L_RECEIPTDATE : date
    - L_SHIPINSTRUCT : fixed text, size 25
    - L_SHIPMODE : fixed text, size 10
    - L_COMMENT : variable text size 44

As mentioned in Section 3.2.3, measuring the memory size in terms of bytes that a row of a table occupies has no single approach, as it depends on how the row is stored (e.g. a DBMS, a row or column-oriented format). Thus no specific figure is given here, but all information on the data from how to retrieve it and how it is composed is provided so that a memory size in bytes can be calculated at need.

Considering the different number of columns of the two tables used, this means that the selected metrics, i.e. latency (seconds) and throughput(rows/second) cannot be used to compare results across different tables.

This is the reason why the comparative evaluation only considers different configurations on the same table.

For this project, five table variations were used to benchmark the code solution as D1. SF was varied to obtain a table at each significant figure, from 10000 rows to 60000000 rows. These are the tables:

1. *supplier_sf1*: size = 10000 rows
2. *supplier_sf10*: size = 100000 rows
3. *supplier_sf100*: size = 1000000 rows
4. *lineitem_sf1*: size = 60000000 rows
5. *lineitem_sf10*: size = 60000000 rows

3.2.5 Experimental Design

The experiments aim to highlight the differences between the newly implemented system based on the delta-rs library, and the current legacy system. To isolate the benefit of using delta-rs over Spark and provide a baseline, three different testing pipelines were designed:

1. **delta-rs - HopsFS**: the system implemented in Chapter 4. It comprises a Rust pipeline with Python bindings, enabling performing operations (i.e., reading, writing) on Delta Lake tables. This pipeline writes on HopsFS.
2. **delta-rs - LocalFS**: this pipeline uses the same library as the system implemented, but saves data on the Local File System (LocalFS). This provides a comparison within the delta-rs library, isolating the impact on performance caused by writing on HopsFS, a distributed file system.
3. **Legacy pipeline**: this pipeline uses the Hopsworks Feature Store to write data on Hudi tables. This system makes use of a pipeline based on Kafka, and Spark to write data on the Hudi tables, saved on HopsFS. The pipeline uses a Spark alternative, namely DuckDB and Arrow Flight, to read data as explained in Section 2.5.2.

Furthermore, the experiments will verify how the performances of the three systems will change based on the CPU resources provided (namely 1 core, 2 cores, 4 cores, 8 cores). Each time the testing environment will be modified accordingly, creating a new VM where the experiments will run with increasingly more resources. These CPU configurations were chosen together with the industrial supervisor, according to the typical Hopsworks use case (Section 3.2.3).

The data used for experiments, as described in Section 3.2.4, will come from two different tables. These are modified according to a SF, for a total of five times.

Additionally, during the writing experiments performed using the legacy pipeline (Spark-based) the contribution of different parts of the process will be measured: namely the upload time and materialization time, dichotomy explained in Section 2.5.1. This will serve to verify how different parts of the legacy pipeline scaled with table sizes, and if Spark was the bottleneck of the architecture.

The Apache Hudi pipelines are preferred over the new Spark based pipeline reading and writing on Delta Lake because these were released and tested extensively on Hopsworks platform, so they provide more guarantees of obtaining relevant results. Additionally, at the time of experiment design these two variations, Spark writing on Delta Lake and Spark writing on Apache Hudi, were considered similarly in read and write performance.

In conclusion, the experiments conducted will be a total of 3 (pipelines) times 4 (CPU configurations) times 5 (tables) times 2 (read and write operations), i.e. 120 experiments, performed 50 times each to create statistically significant results.

3.2.6 Experimental environment

The experimental environment consists of a physical machine in Hopsworks' offices, virtualized enabling remote shared development. The CPU details of the machine are present in Listing 3.1, noting that only eight cores at maximum were accessed during the experiments. It should be observed that this experimental environment while virtualized in isolation, runs on shared computing resources, so experiment results might vary depending on the load of the machine. Considering this all experiments will be run when the machine load is low (less than 50% of CPU and RAM usage), to avoid having results depend on external workloads running.

The machine mounts about 5.4 TBs of SSD memory. This allows the machine to have fast read and write speed, 2.7 GB/s, and 1.9 GB/s respectively (measured with a simple *dd* bash command).

The experimental environment will be set up with a Jupyter Server of different CPU cores, depending on the experiment (1 core, 2 cores, 4 cores, or

8 cores). The Jupyter server is allocated by default with 2048 MB of RAM, out of the 192 GB available on the experimental machine. This amount will be adjusted during the experiments according to the needs of the experiments (see Section 5.1.4).

Listing 3.1: Output of a *lscpu* bash command on the test environment.

```
Architecture:                x86_64
  CPU op-mode(s):            32-bit, 64-bit
  Address sizes:             48 bits physical,
                             48 bits virtual
  Byte Order:                Little Endian
CPU(s):                      32
  On-line CPU(s) list:       0-31
Vendor ID:                   AuthenticAMD
  Model name:                AMD Ryzen Threadripper
                             PRO 5955WX 16-Cores
    CPU family:              25
    Model:                   8
    Thread(s) per core:      2
    Core(s) per socket:      16
    Socket(s):               1
    Stepping:                2
    Frequency boost:         enabled
    CPU max MHz:             7031.2500
    CPU min MHz:             1800.0000
    BogoMIPS:                7985.56
Virtualization features:
  Virtualization:            AMD-V
Caches (sum of all):
  L1d:                       512 KiB (16 instances)
  L1i:                       512 KiB (16 instances)
  L2:                        8 MiB (16 instances)
  L3:                        64 MiB (2 instances)
```

3.2.7 Evaluation Framework

The system evaluation framework is designed to evaluate three key aspects of the system, using different metrics:

1. **Functional requirements**: defined in Section 3.1.3, functional requirements will be measured by verifying the success or failure of running an experiment. By design, this will not happen, as the system implementation phase, continues until all functional requirements are met.

2. **Non-functional requirements**: defined in Section 3.1.3, non-functional requirements are: consistency, maintainability and scalability. The first two requirements are mainly addressed during implementation, while scalability is measured during the system evaluation experiments. The metric used for measuring this requirement is the throughput measured in rows/second as defined in RQ2.

3. **How does the system compare to other pipelines?**: this question answers directly RQ2, measuring the throughput (rows/second) of the different pipelines, defined in Section 3.2.5. Results are then compared using a visual approach.

3.2.8 Assessing Reliability and Validity

Results are significant according to their reliability and validity. In this project work, to ensure the reliability of the experiment results on the system performance, each experiment will be performed fifty times. This number was agreed as a balance between consistency and the limited resources available (in terms of time and computing resources).

Probably due to the complex nature of the pipeline tested, the data distribution of results could vary from one experiment to the other. This hampers the possibility of comparing results, greatly impacting the relevance of the results analysis. To restore the validity of the data collected a bootstrapping technique will be used. Data will be resampled with substitution a thousand times, then an average with a confidence interval for each experiment will be calculated. This will benefit the accuracy of the results presented.

Chapter 4

Implementation

This chapter follows the system implementation process defined in Section 3.1 detailing and explaining how the system was developed, how the deploy it, use it and how the code implemented was run during the experimental part of this thesis work.

4.1 Software design and development

The first step of the development process, defined in Section 3.1.2, consists of identifying what the delta-rs library (version 0.15.x) lacks to satisfy the requirements, more specifically, how to support HDFS and HopsFS in the delta-rs [29] library. This step outlines the library structure (colloquially defined as "crate") divided into sublibraries (colloquially defined as "subcrates"), which is illustrated in Figure 4.1. As the figure shows, the delta-rs crate has a subcrate for each storage connector, so adding support for different storage is a matter of implementing a new subcrate to the delta-rs crate.

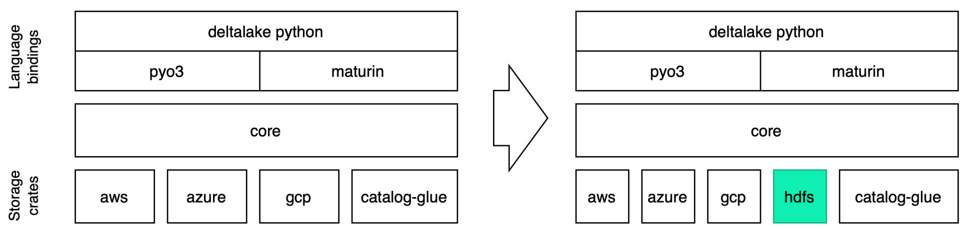

Figure 4.1: delta-rs library (v. 0.15.x) before and after adding HDFS support

The delta-rs library uses an external interface called object-store defined in the Arrow-rs library [67]. Every storage connector implements this interface, and the other parts of the library interact with the storage layer. It is thus crucial that the new HDFS storage also implements this interface.

The development process was divided into two subsections: a first approach and a final approach. The first approach was carried out by developing a HDFS subcrate for the delta-rs crate from scratch, while the second and final approach modified the support for HDFS added in delta-rs with version 0.18.2 to also support HopsFS.

4.1.1 First approach

At the beginning of the project, a first approach was taken to provide support for HDFS to the delta-rs library version 0.15.x. Analyzing past contributions to the library reveals that HDFS was supported in version 0.9.0 of the delta-rs library, but support was removed due to three main reasons:

1. The HDFS support caused the testing pipeline to fail, and no trivial solution was found.

2. The HDFS support had JVM dependencies. This was considered a strong limitation for Python users as having Java installed is high overhead (in terms of storage and computation) for performing an operation in Python. Having these Jave dependencies would have meant that a large number of Python users would have not used the library altogether.

3. The community around the library did not have contributors or a large number of users which had HDFS as a use case.

Starting from the HDFS support in the delta-rs library present in version 0.9.0, a solution was designed to fix the testing issues and provide a working storage support for HDFS. The architecture is described in Figure 4.2.

This implementation makes use of a C++ library called libhdfs, which contains all the methods required to work as an HDFS client. This library is contained in the Rust library fs-hdfs. The datafusion-objectstore-hdfs makes use of the libhdfs library, to provide an interface for HDFS that implements object-store, the interface used in the delta-rs library to interact with storage connectors.

This approach required first to rewrite the datafusion-objectstore-hdfs as it made use of an older object-store interface version (0.8.0 vs. 0.10.0) and it was

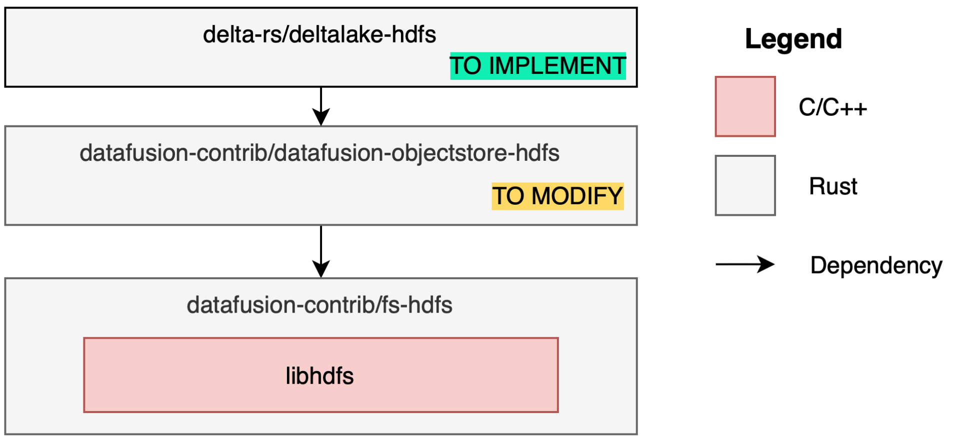

Figure 4.2: Architecture of the first implementation approach

not possible anymore to upgrade it easily. Secondly, the JVM dependencies while this project did not have a strict requirement on not having them, being able to have the dependencies consistently work on different development environments proved to be a challenge.

Ultimately, this approach was abandoned following the release of version 0.18.2 of the delta-rs library. This decision was taken to comply with the maintainability requirement defined in Section 3.1.3.

4.1.2 Final solution

Version 0.18.2 of the delta-rs introduced support for HDFS via the hdfs-native library [68]. This is a Rust library that re-implements the HDFS client, avoiding to use of libhdfs, and thus has no JVM dependencies. This architecture is illustrated in Figure 4.3.

This section approach, while being used by the delta-rs library, proved to have some incompatibilities with HopsFS. Here below is a list of them:

1. **Different HDFS protocol version**: HDFS makes use of a Remote Procedural Call (RPC) protocol to interact with a HDFS client. Hdfs-native is based on protocol version 3.2, while HopsFS was compatible with version 2.7.

2. **No support for TLS in hdfs-native**: hdfs-native security is based on Kerberos, but it does not secure packet transfers using TLS.

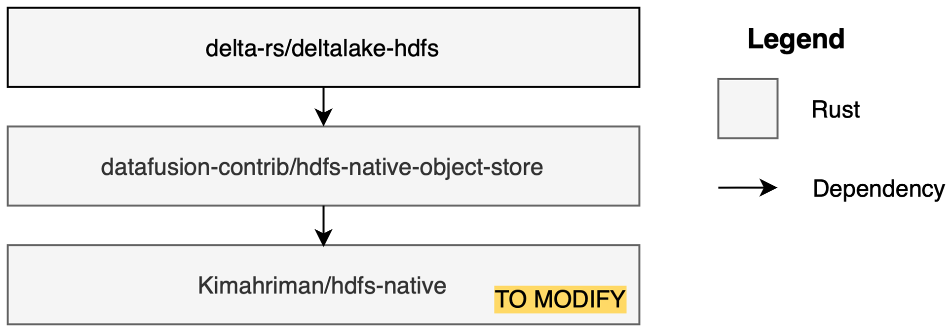

Figure 4.3: Architecture of the final implementation approach

These incompatibilities were solved one by one during development in the following way:

1. **Upgrading HopsFS protocol version**: together with the industrial supervisor, responsible for the maintenance of HopsFS, the differences between the two protocol versions (2.7 vs. 3.2) were highlighted, and one by one resolved. This way version 3.2.0.14 of HopsFS was released, adding support to the HDFS RPC protocol version 3.2, making HopsFS compatible with the hdfs-native library.

2. **Adding support for TLS in the hdfs-native library**: TLS support to hdfs-native was added via the use of an external Rust library called tokio-rustls version 0.26 [69].

All changes were applied to copies (colloquially defined as "fork") of the open-source repositories related to this project (delta-rs, hdfs-native-object-store, hdfs-native). These changes are available on the author's repository [70, 71, 72].

4.2 Software deployment and usage

Once HDFS and HopsFS support has been added to the delta-rs library it is sufficient to build a python wheel, i.e. a pre-built binary package format for Python modules and libraries, for delta-rs. To do so it is sufficient to follow the instructions already present in the delta-rs library in the README.md file present in the python folder [73].

The usage of the delta-rs library is explained in detail in the delta-rs documentation [29], so in this section, only the method used for the experiments will be shown and explained. Listing 4.1 shows a simple example of writing to HDFS or HopsFS (formally in a Python script only the address changes, as HopsFS is based on HDFS). As is clear from the listing in this case the script is writing a small table of two columns and three rows.

Listing 4.1: Writing a data frame on a Delta Table with delta-rs on HDFS or HopsFS

```
from deltalake import write_deltalake
import pandas as pd

df = pd.DataFrame({"num": [1, 2, 3],
                   "letter": ["a", "b", "c"]})
write_deltalake("hdfs://rpc.sys:8020/tmp/test", df)
```

In Listing 4.2 an example of a read operation is shown. After being read, the Delta Table is converted to a pyarrow table, to ensure an explicit in-memory operation (only calling the DeltaTable object is a lazy operation that does not load the data into memory).

Listing 4.2: Reading a data frame on a Delta Table with delta-rs on HDFS or HopsFS

```
from deltalake import DeltaTable

dt = DeltaTable("hdfs://rpc.sys:8020/tmp/test")
dt.to_pyarrow_table()
```

4.3 Experiments set-up

Experiments, as defined in Section 3.2.5, consist of running different system configurations, with different data, fifty times per experiment. Two main approaches were selected to measure the experiment's time, here is explained how to set them up.

The first approach is to use the Python timeit function. As illustrated in Listing 4.3 timeit can be used by defining a SETUP_CODE that runs before the experiment and a TEST_CODE that when running is measured and the time (expressed in seconds) is the return value of the timeit function. This approach was selected as the timeit function provides a clear interface to run and measure a small code script. The script here does not run a repeated number of times as the Delta Table must be deleted before re-running the experiment, and this requires time that shouldn't be included in the experiment time (nor in the setup, as the first time the table is not there). When this approach could not be used (due to more complex scripts, the second approach was used).

Listing 4.3: Timeit usage to measure the time required to write a Delta Lake table to HopsFS.

```
import timeit
SETUP_CODE='''import pyarrow as pa
from deltalake import write_deltalake'''

TEST_CODE='''
HDFS_DATA_PATH = "hdfs://rpc.sys:8020/exp"
LOCAL_PATH = "/abs/path/table.parquet"
pa_table = pa.parquet.read_table(LOCAL_PATH)
write_deltalake(HDFS_DATA_PATH, pa_table)'''

# Measure the execution runtime
write_result = timeit.timeit(setup  = SETUP_CODE,
                             stmt   = TEST_CODE,
                             number = 1          )
```

The second measuring approach was to simply record the time before the script run and after the script run, then calculate the difference. This made it possible to calculate multiple differences without having to recreate the experiment multiple times. Listing 4.4 shows an example of this approach.

Listing 4.4: A simple time difference approach to measure the time required to write a Delta Lake table to HopsFS.

```
import time
import pyarrow as pa
from deltalake import write_deltalake
HDFS_DATA_PATH = "hdfs://rpc.sys:8020/exp"
LOCAL_PATH = "/abs/path/table.parquet"
pa_table = pa.parquet.read_table(LOCAL_PATH)

before_writing = time.time()
write_deltalake(HDFS_DATA_PATH, pa_table)
after_writing = time.time()
```

```
write_result = after_writing - before_writing
```

Chapter 5

Results and Analysis

This chapter is the output of the system evaluation process defined in Section 3.2. It starts with Section 5.1, which presents the results of the experiments performed in the form of tables, histograms and written descriptions. Then Section 5.2 complements the chapter by analyzing and discussing the results' findings.

5.1 Major Results

This section presents the main results of the 120 experiments performed as defined in Section 3.2.5. The experiments are grouped into subsections according to the measured operation (read or write). In each one of the subsections histograms and tables are present to visualize the results using both metrics (latency expressed in seconds and throughput expressed in rows/second).

Results are reported using the log scale for clarity, as results differing from more than one significant figure are not clearly visible using a histogram representation. Additionally, for each measurement displayed a 95% confidence interval was also calculated using the bootstrapping technique mentioned in Section 3.2.8. Nonetheless, this interval was not reported in the histograms in this section, as it would be hardly readable as all results are out of each other's 95% confidence interval. It can still be visioned in the Appendix

Add ref to appendix

where for each experiment a histogram and a table containing the 95% confidence interval were reported.

Considering that latency and throughput are inversely correlated (see equation in Section 3.2.2) trends observed when measuring latency are inversely reflected when data is observed as throughput (e.g. if latency is halved, throughput doubles, if latency quarters, throughput quadruples). This is because all experiments were performed with fixed-size tables.

Due to this correlation between the metrics used, trends will be described using latency (the measured metric), and complemented in brackets for throughput (the computed metric) using an abbreviation (thr:).

5.1.1 Writing Experiments

Fix this ref to non-existing table+figure

Figure **??** and Figure **??** show the write latency and throughput respectively of the write operations performed on three different systems defined in Section 3.2.5 when writing the five different tables defined in Section 3.2.4. Both histograms, i.e. Figures 5.1 and 5.2, report the data from the 1 CPU core experiment while the tables, i.e. Tables 5.1 and 5.2, report both the 1 CPU core experiment data and also a calculated percentage of improvement (decrease in the case of latency, increase in the case of throughput) of the specified metric as the CPU cores increase.

delta-rs on HopsFS vs. delta-rs on LocalFS

The latency (thr: throughput) measured using the delta-rs on LocalFS pipeline results around ten times lower (thr: higher) than the latency (thr: throughput) measured using the delta-rs on HopsFS pipeline for small tables (10K and 100K rows). On the other hand, the latency (thr: throughput) in the two pipelines is more similar (same significant figure) on experiments performed with larger tables (1M, 10M, 6M, and 60M rows). Overall the latency (thr: throughput) measured in the delta-rs on LocalFS pipeline remains lower (thr: higher) in absolute terms than the latency (thr: throughput) measured using the delta-rs on HopsFS pipeline in all experiments.

delta-rs on HopsFS vs. Legacy pipeline

The latency (thr: throughput) measured using the delta-rs on HopsFS pipeline results more than ten times lower (thr: higher) than the latency (thr: throughput) measured using the Legacy pipeline in all experiments. This trend results more prominent for smaller tables (10K and 100K rows) where latency (thr: throughput) measured using the delta-rs on HopsFS pipeline is forty

Table 5.1: Write latency experiment results compared across multiple CPU configurations

Pipeline	Number of rows	1 CPU core latency (seconds)	2 CPU cores (% decrease)	4 CPU cores (% decrease)	8 CPU cores (% decrease)
delta-rs HopsFS	10K	1.250 88	−0.92	2.75	−9.33
	100K	1.368 28	4.40	2.34	5.54
	1M	9.381 52	9.23	10.32	11.52
	6M	19.754 69	17.54	17.87	20.33
	60M	177.307 07	24.39	30.01	31.22
delta-rs LocalFS	10K	0.039 57	−21.88	−15.53	−11.25
	100K	0.152 40	10.01	13.54	10.45
	1M	8.422 52	14.69	14.68	14.17
	6M	17.906 34	14.74	18.71	20.24
	60M	172.345 52	24.67	29.57	30.38
Legacy	10K	50.227 67	−0.99	−2.10	−1.99
	100K	59.561 87	−0.38	0.06	−1.20
	1M	112.190 48	3.23	3.01	2.50
	6M	511.816 93	7.51	5.83	7.01
	60M	2 715.772 85	13.81	13.61	14.39

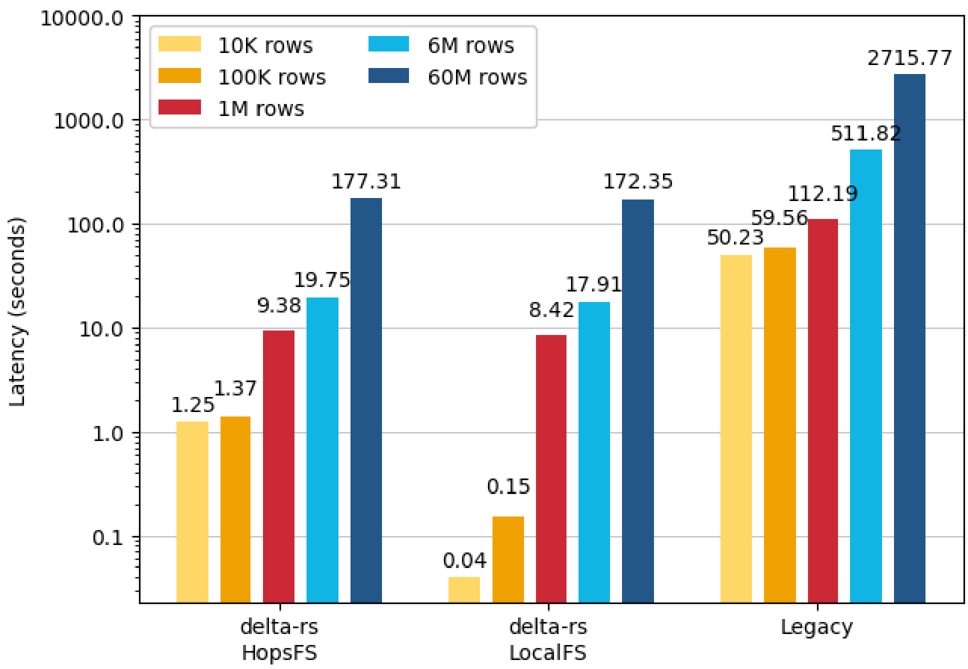

Figure 5.1: Histogram in log-scale of the write latency experiment results performed with 1 CPU core

Table 5.2: Write throughput experiment results compared across multiple CPU configurations

Pipeline	Number of rows	1 CPU core throughput (k rows/second)	2 CPU cores (% increase)	4 CPU cores (% increase)	8 CPU cores (% increase)
delta-rs HopsFS	10K	7.994 36	−0.91	2.83	−8.53
	100K	73.084 38	4.60	2.40	5.87
	1M	106.592 42	10.17	11.51	13.01
	6M	303.725 33	21.27	21.76	25.52
	60M	338.395 98	32.26	42.87	45.39
delta-rs LocalFS	10K	252.682 38	−17.95	−13.44	−10.12
	100K	656.157 39	11.13	15.66	11.67
	1M	118.729 19	17.22	17.21	16.51
	6M	335.076 75	17.29	23.02	25.38
	60M	348.137 84	32.76	41.99	43.65
Legacy	10K	0.199 09	−0.98	−2.06	−1.95
	100K	1.678 92	−0.38	0.06	−1.19
	1M	8.913 41	3.34	3.10	2.57
	6M	11.722 94	8.12	6.19	7.54
	60M	22.093 15	16.02	15.76	16.81

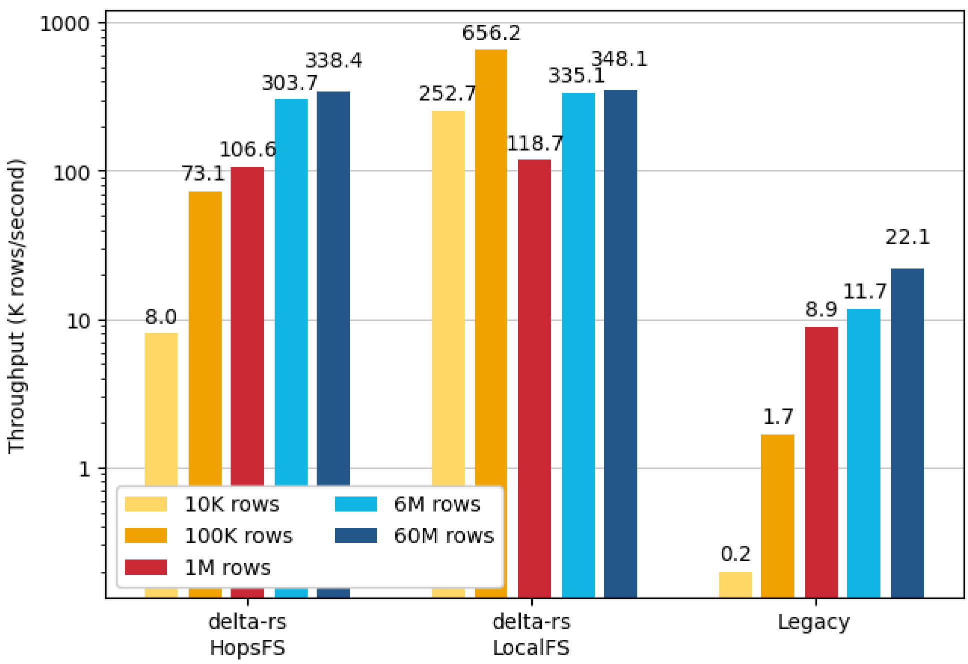

Figure 5.2: Histogram in log-scale of the write throughput experiment results performed with 1 CPU core

times lower (thr: higher) than the latency measured using the Legacy pipeline. While the tendency is less marked for larger tables (6M and 60M rows) where latency (thr: throughput) measured using the delta-rs on HopsFS pipeline is around twenty times lower (thr: higher) than the latency (thr: throughput) measured using the Legacy pipeline.

Change of performance as the CPU cores increase

During experiments with more CPU cores, in delta-rs based pipelines (writing on HopsFS or LocalFS) the latency (thr: throughput) during write operation decreases (thr: increases) by a considerable amount: 20-30% (thr: 30-40%), only when writing larger tables (6M and 60M rows), while it decreases (thr: increases) by a lower margin: 5-10% (thr: 5-15%) on smaller tables (100K and 1M rows), even slightly increasing (thr: decreasing) on the smallest table (10K rows). It should be noted that latency (thr: throughput) decreases (thr: increases) as described in the 2 CPU cores experiments, remaining on similar improvements even with more CPU cores.

Considering the Legacy pipeline, experiments with more CPU cores did not decrease (thr: increase) the latency (thr: throughput) by more than 7% (thr: 8%) except for the largest table (60M rows). This table benefitted from a latency (thr: throughput) decrease (the: increase) of around 14% (thr: 16%). The smallest tables (10K and 100K) even reported slight increases (thr: decreases) in the latency measured.

5.1.2 Reading Experiments

Fix these refs

Figure **??** and Figure **??** show the read latency and throughput respectively of the read operations performed on three different systems defined in Section 3.2.5 when reading the five different tables defined in Section 3.2.4. Both histograms, i.e. Figures 5.3 and 5.4, report the data from the 1 CPU core experiment while the tables, i.e. Figures 5.3 and 5.4, report both the 1 CPU core experiment data and also a calculated percentage of improvement (decrease in the case of latency, increase in the case of throughput) of the specified metric as the CPU cores increase.

delta-rs on HopsFS vs. delta-rs on LocalFS

The latency (thr: throughput) measured using the delta-rs on LocalFS pipeline results around ten times lower (thr: higher) than the latency (thr: throughput)

Table 5.3: Read latency experiment results compared across multiple CPU configurations

Pipeline	Number of rows	1 CPU core latency (seconds)	2 CPU cores (% decrease)	4 CPU cores (% decrease)	8 CPU cores (% decrease)
delta-rs HopsFS	10K	0.053 42	22.65	18.84	18.95
	100K	0.057 57	1.15	3.76	5.19
	1M	0.538 55	56.53	65.00	67.71
	6M	1.948 99	53.40	72.74	74.48
	60M	22.980 65	50.34	75.72	87.20
delta-rs LocalFS	10K	0.004 19	31.48	35.91	29.66
	100K	0.026 96	51.54	65.76	64.84
	1M	0.420 09	52.45	78.64	89.75
	6M	1.682 23	55.56	77.99	89.57
	60M	19.565 47	51.72	75.41	88.32
Legacy	10K	0.631 59	1.06	−0.67	0.67
	100K	2.650 10	−0.50	0.39	−0.46
	1M	8.596 36	−0.24	−1.81	2.89
	6M	33.529 64	0.46	0.23	0.30
	60M	33.697 72	0.16	0.13	1.64

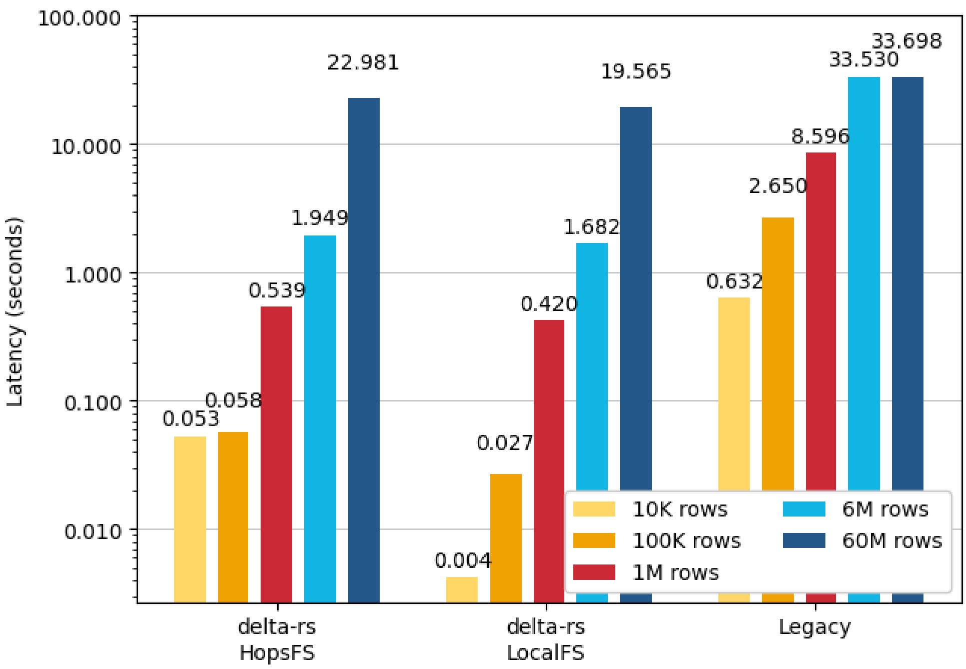

Figure 5.3: Histogram in log-scale of the read latency experiment results performed with 1 CPU core

Table 5.4: Read throughput experiment results compared across multiple CPU configurations

Pipeline	Number of rows	1 CPU core throughput (k rows/second)	2 CPU cores (% increase)	4 CPU cores (% increase)	8 CPU cores (% increase)
delta-rs HopsFS	10K	187.168 53	29.28	23.21	23.38
	100K	1 736.907 99	1.17	3.90	5.48
	1M	1 856.831 67	130.02	185.74	209.69
	6M	3 078.512 99	114.57	266.87	291.92
	60M	2 610.891 46	101.35	311.83	681.03
delta-rs LocalFS	10K	2 384.586 99	45.94	56.04	42.18
	100K	3 708.257 87	106.37	192.07	184.38
	1M	2 380.403 81	110.28	368.24	875.15
	6M	3 566.674 54	125.01	354.40	858.64
	60M	3 066.626 44	107.11	306.75	756.07
Legacy	10K	15.832 85	1.07	−0.67	0.67
	100K	37.734 32	−0.50	0.39	−0.45
	1M	116.328 20	−0.24	−1.78	2.98
	6M	178.946 12	0.46	0.23	0.30
	60M	1 780.535 63	0.16	0.13	1.67

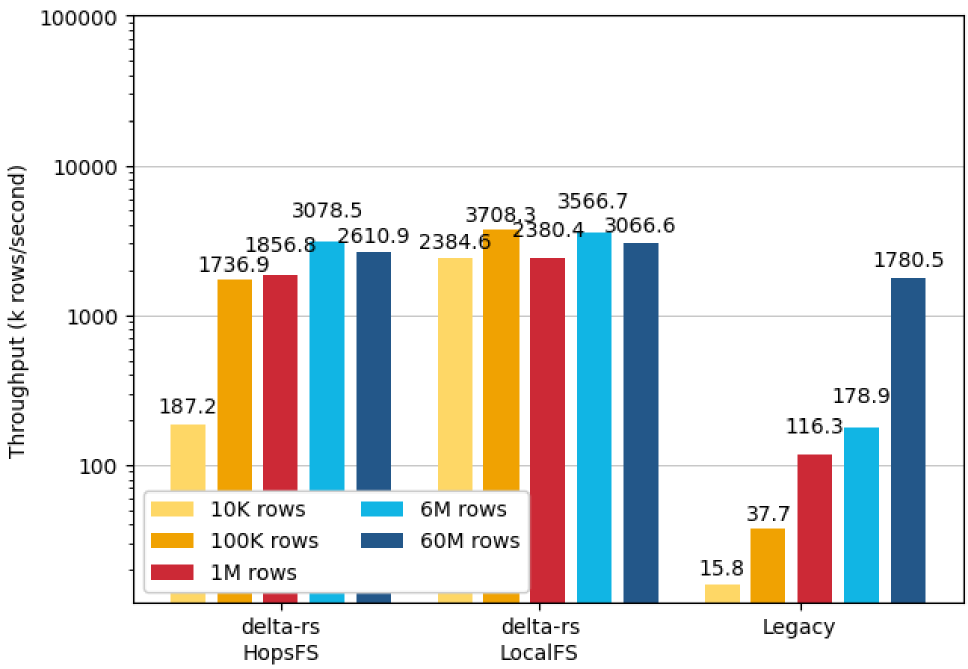

Figure 5.4: Histogram in log-scale of the read throughput experiment results performed with 1 CPU core

measured using the delta-rs on HopsFS pipeline in the experiment with the smallest table (10K rows). On the other hand, the latency (thr: throughput) in the two pipelines is more similar (same significant figure) on experiments performed with larger tables (100K, 1M, 10M, 6M, and 60M rows). Overall the latency (thr: throughput) measured in the delta-rs on LocalFS pipeline remains lower (thr: higher) in absolute terms than the latency (thr: throughput) measured using the delta-rs on HopsFS pipeline in all experiments.

delta-rs on HopsFS vs. Legacy pipeline

The latency (thr: throughput) measured using the delta-rs on HopsFS pipeline results more than ten times lower (thr: higher) than the latency (thr: throughput) measured using the Legacy pipeline in all but the experiment with the largest table (60M rows), where the latency (thr: throughput)of the first pipeline is only 47% lower (thr: higher) than the second.

Change of performance as the CPU cores increase

During experiments with more CPU cores, in delta-rs based pipelines (reading on HopsFS or LocalFS) the latency (thr: throughput) during read operation decreases (thr: increases) by a considerable amount: +50% (thr: +100%), when reading larger tables (1M, 6M and 60M rows), while it decreases (thr: increases) by a lower margin: 20-30% (thr: 30-45%) on smaller tables (10K and 100K rows). It should be noted that latency decreases (thr: increases) in reads with larger tables (1M, 6M, and 60M rows) following an inverse linear relationship with the increase of CPU cores: 2 CPU cores, latency halves, 4 CPU cores, latency quarters, 8 CPU cores, latency is decreased to an eighth. On the other hand, throughput follows a linear relationship with the increase of CPU cores.

Considering the Legacy pipeline, experiments with more CPU cores did not decrease (thr: increase) the latency (thr: throughput) by more than 2% (thr: 8%). Histograms comparing the three pipelines, look radically different in experiments with more CPU cores, due to how delta-rs scales with the increase of CPU cores. They can be accessed in the Appendix

Add ref. to appendix

.

Table 5.5: Contributions to the write latency of the upload and materialization steps in the legacy pipeline and its change at different CPU configurations

Number of rows	1 CPU core latency (seconds)		2 CPU cores (% decrease)		4 CPU cores (% decrease)		8 CPU cores (% decrease)	
	upl.	mat.	upl.	mat.	upl.	mat.	upl.	mat.
10K	2.48	47.72	3.99	−1.27	4.10	−2.47	4.22	−2.35
100K	3.66	55.90	6.37	−0.78	5.55	−0.27	6.25	−1.69
1M	22.59	89.57	17.52	−0.39	14.89	−0.01	16.51	−1.05
6M	244.61	267.24	15.83	−0.10	13.46	−1.15	15.10	−0.38
60M	2 437.78	278.05	15.33	0.39	15.15	0.16	15.94	0.82

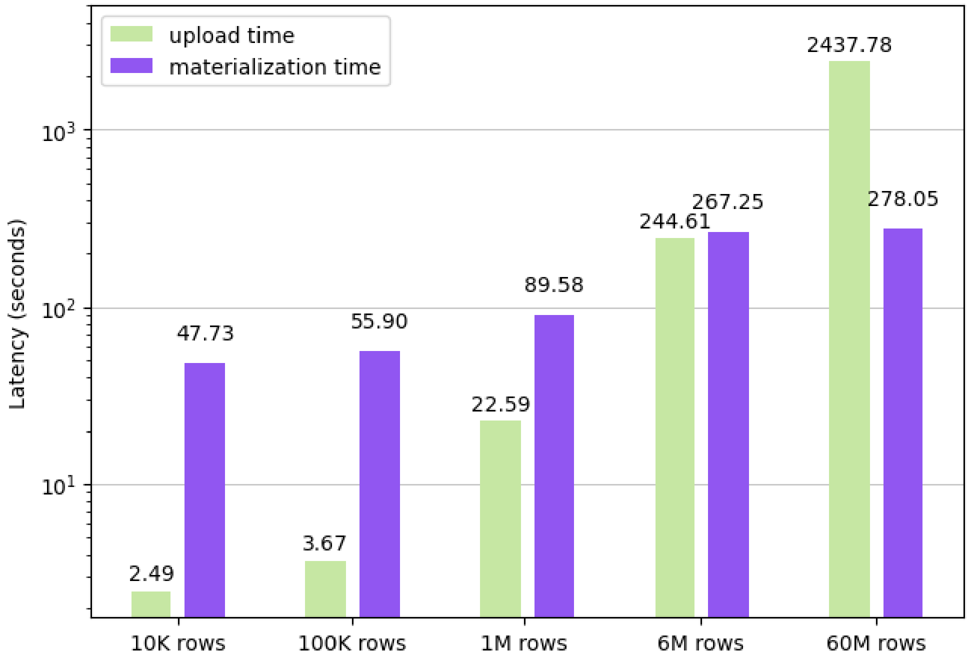

Figure 5.5: Histogram in log-scale displaying the contributions to the write latency of the upload and materialization steps in the legacy pipeline

5.1.3 Legacy pipeline write latency breakdown

Fix ref

Figure **??** shows the write latency breakdown of the Legacy pipeline into upload time and materialization time, the different steps of the Legacy pipeline as explained in Section 2.5.1. The breakdown is proposed for all the five different tables defined in Section 3.2.4. Figure 5.5 reports the data from the 1 CPU core experiment while Figure 5.5 reports both the 1 CPU core experiment data and also a calculated percentage of improvement (decrease) of the latency as the CPU cores increase.

Considering the upload time contribution to the write latency, this represents a small percentage (around 5%) when writing smaller tables (10K and 100K rows), while its contribution grows following a similarly linear pattern in larger tables (between 100K and 60M rows). This changes radically the proportion between the upload and materialized contribution to the write latency, making the upload time 90% of the total write latency for the 60M rows table. On the other hand, the materialization time contribution while starting with high latency contribution (95% of the total), its absolute value does not increase by a considerable amount (less than a significant figure) even if the table size increased by three significant figures.

Observing the results of experiments using more CPU cores, the upload time benefits from a higher number of CPU cores in particular with larger tables (15% latency decrease) and less with smaller tables (4% latency decrease). On the other hand, the materialize time does not improve performance having either small decreases in latency or small increases (both around 1-2%).

5.1.4 In-memory resources usage

The experimental environment resources defined in Section 3.2.6 were adjusted according to the computational needs. In particular, write operations were demanding on the available RAM resources, requiring up to 24 GBs to operate with the larger tables (6M and 60M rows). The system was adjusted to allocate 32768 MB of RAM, so it could avoid slowing down operations.

5.2 Results Analysis and Discussion

This section discusses the main results presented in Section 5.1, trying to explain their meaning, their implication for the company, and more generally for the research area. Additionally, this section provides a collection of project considerations outlined either during development or after the experiments conducted.

5.2.1 Discussion on main results

The results presented in Section 5.1 reveal the differences in latency (and thus also data throughput) between the newly implemented system using the delta-rs library and the legacy system. While the experiments present task-specific differences in performance between the two systems, overall the system developed in this thesis work has at least ten times lower latency than the legacy system in all experiments when using the tables from 10K to 6M rows. These findings not only confirm the hypothesis that a Rust-based system would be faster than a Spark-based system when operating on small tables (from 10K to 6M rows) but also show that delta-rs is a preferable alternative to how read operations are currently performed using a combination of Arrow Flight and DuckDB (Section 2.5.2). Said outcomes further solidify the idea of the pivotal role that Rust will play in optimizing computer systems [74].

Taking a more in-depth look at the results of the writing experiments, something more can be said. In writing experiments, even with the largest table (60M rows), the newly implemented system has a latency ten times lower than the legacy system. When considering the smallest tables (10K and 100K rows), the improvement of the new system over the legacy pipeline is up to forty times. This confirms even more the need for Spark alternatives when elaborating small-size data. For the use case defined in Section 3.2.3 it is clear that the new system is a better alternative in terms of performance but also costs, as maintaining a Spark cluster for this small amount of data makes little sense.

Moving on to reading experiments, as explained in Section 2.5.2, the legacy system when reading is already using a Spark alternative, i.e. Arrow Flight and DuckDB. The results show that there is a smaller difference between the two systems when reading the largest table (60M rows) with only 46% improvement. Nonetheless, the new system scales much better as the CPU cores increase, up to 89.75% with eight cores, while performance in the legacy system remains more or less the same. This has to do with how resources are allocated in the system. The user can modify dynamically his local resources, and so since delta-rs uses local resources to operate the user can tune the system at his necessity. On the other hand, the Arrow Flight server using DuckDB has a predefined amount of resources available, that cannot be modified as easily (the Hopsworks cluster needs to be redeployed). Overall the flexibility that the new system offers, at the expense of delegating computation on the client side is remarkable and might benefit hybrid workflows.

The impact of these findings for Hopsworks AB is considerable, as it affects their main product, i.e. the Hopsworks Feature Store. The results recommend adopting the developed system over the current system using Apache Hudi as an Offline Feature Store. Considering the Hopsworks AB approach of supporting multiple ways to load and save data, an alternative to the total substitution of the system would be leaving the option to the user to choose which data lakehouse format the Offline Feature Store should save the data. It should be noted that the experiments and system evaluation were performed on the use case defined in Section 3.2.3. For different use cases, more experiments should be performed. The author expects that there will be a data size threshold where using a Spark-based will make more sense in terms of performance (lower latency).

More generally speaking, these findings have little possibility of being generalized due to the intrinsic bias of conducting an industrial master thesis within the company developing the product to evaluate. Furthermore, the defined use case (Section 3.2.3) restricts the results on specific table sizes and computational resources. Results could vary if reading or writing more data with a different amount of resources. Nonetheless, the results confirm the research expectations, contributing to supporting the idea that Spark alternatives should be considered (and sometimes preferred) when working with small-sized tables (100K to 60M rows). This encourages more research and experiments to be conducted on Spark alternatives and Rust applications in data management system implementations.

5.2.2 Considerations on the legacy system

The legacy system presented in Sections 2.5.1 and 2.5.2 has a layered architecture design to fit multiple needs. For example, Kafka is used as a single point of upload of data both to the Offline Feature Store analyzed in this thesis work and the Online Feature Store designed for streaming pipelines. Using Kafka ensures that data is consistent between the two Feature Stores (Online and Offline) but it also represents a limitation in the system performance. As results in Figure **??** show, Kafka as data increases represents the bottleneck of the architecture, making 95% of the write latency when writing a 60M rows table. This has mainly to do with how Kafka is used in the architecture and how data is sent, i.e. row by row. This work helped highlight this issue, but more research and experiments are required to find an effective solution. Enabling multiple uploads via concurrency mechanisms might speed up the upload process. Alternatively, sending columns instead of rows, and keeping a column structure along the pipeline might also help.

Another aspect that limited the capability of the legacy architecture is how resources are allocated for the Spark cluster. These can be modified more accessibly compared to the computation resources for the Arrow Flight server, but having to balance two systems (the client's and the Spark client) creates an added complexity not necessary when elaborating small quantities of data.

5.2.3 Considerations on the delta-rs library

The system implementation focus of this project was adding support for HopsFS to the delta-rs library. This was made possible also thanks to the built-in modularity that the library offers, having a different sub-library for each storage connector. The community around the library is very active and each question made on their communication channels, namely GitHub and Slack, would always receive an answer within a few hours or a day. The only recommendation the author would give to the library's maintainers would be to document further, perhaps with the help of architectural diagrams the inner workings of the library's processes, specifying how and when data gets uploaded into memory. The first iteration of the reading experiments had to be scratched due to calling a lazy function that would not load data into memory.

Considering the results presented in Section 5.1 the similarities in performance (latency and throughput) of the two delta-rs pipelines, operating on the LocalFS and HopsFS respectively suggests that the library is operating at its full potential also in the newly implemented system. Delta-rs on the LocalFS still performs faster, but this might just have to do with the different nature of the two storage systems, i.e. local and distributed.

Chapter 6

Conclusions and Future work

This chapter presents the conclusions of the thesis work. It starts with Section 6.1 that compares the experimental findings with the Research Question to outline this thesis contribution. Section 6.2 then explains the limitations of the project in terms of resources and scope. Finally, Section 6.3 complements the chapter by considering possible future work stemming from this thesis.

6.1 Conclusions

This thesis work posed two RQs defined in Section 1.2.1. These were:

RQ1: How can we add support for HDFS and HopsFS to the delta-rs library?

RQ2: What is the difference in latency and throughput between the current legacy system (Spark-based in writing) reading and writing to Apache Hudi compared to a delta-rs library-based reading and writing to Delta Lake, in HopsFS?

The work conducted in this thesis answered these two questions, by performing a system implementation and then evaluating the newly implemented system.

The first step was adding support for HopsFS in the delta-rs library. This was achieved by modifying the hdfs-native [68] library, which reimplements in Rust a HDFS client.

The second step was measuring the newly implemented system's performance and comparing it with the legacy system. The metrics used were the latency (seconds) of the read and write operations and the throughput (rows/second) which was calculated by dividing the table size (in rows) by the latency of the operation. The results presented in Chapter 5 revealed that the delta-rs-based access to Delta Lake has a latency at least ten times lower in both read and write operations for tables from 10K to 6M rows in size. Considering only the write experiments these show that the delta-rs library performs up to forty times better with smaller tables (10K and 100K), while still outperforming by ten times the legacy pipeline on the largest table (60M rows). This difference suggests that with larger tables there might be a threshold where a Spark-based system would perform better, but more experiments with larger tables are needed to verify the trend. Similarly in the reading experiments delta-rs outperforms the legacy pipeline more with smaller tables (10K and 100K rows), even if only by a factor of fifteen instead of forty seen for the writing experiments. This is probably caused by the use of a Spark alternative in the reading process of the legacy system (Arrow Flight and DuckDB), which already improves Spark performances on smaller tables. One last notable finding on the difference between the newly implemented system and the legacy pipeline is the difference in scalability as resources increase. Experiments were conducted with an increasing number of CPU cores (from 1 to 8 CPU cores) and the results showed that delta-rs, being a local process, is much more suited for making use of those resources with an up to 31% reduction in latency during writing and a 87% reduction during reading.

Overall, the experiments' results recommend the adoption of the newly implemented system in the defined use case (Section 3.2.3), either in substitution or as an alternative to provide the user if they want to save the data on a Delta Table in the Offline Feature Store.

6.2 Limitations

The limitations of this study mostly derive from the constraints of resources, in terms of time and computational resources, and scope, which is mostly linked to the defined use case (Section 3.2.3).

Being the scope of this project improving the latency of the Hopsworks Offline Feature Store, this is the technology on which the implementation and the experiments were based. This outlines the limited generalization of the results obtained, which are biased from the use of technology in collaboration with the company developing it. Additionally, a specific use case was defined (Section 3.2.3) to choose the CPU loads and data loads on which the experiments would be conducted. This helps to define a perimeter of the thesis contribution but also limits the thesis impact on this specific use case, requiring more research to verify the same hypothesis in a different scenario.

The computational resources provided, while great in size, were used on a shared environment that could only be used for a limited amount of time and only if it was not operating other, more critical workloads. Time also played a role in limiting the number of experiments to be conducted to calculate a 95% confidence interval, all experiments were run 50 times, which increases by a great factor the time required to perform all experiments.

6.3 Future work

The results and limitations of this thesis offer a good starting point for future work. As outlined in the limitations section before, this thesis's scope and resources were limited so conducting new experiments on the system performance by relaxing one or more constraints, could bring new results that can be more general.

Considering the system research contribution of this thesis, this could be expanded mostly for Hopsworks AB needs as the code works with key components of their infrastructure, e.g. HopsFS, that while open-source does not see much use outside the company. The code is unlikely to see use in the delta-rs library, as even if this is a solid contribution, it only fits a single company's use case, and it is very limited in their applications outside it. Future system contributions could still use the code developed in this thesis as a baseline, to be compared with new delta-rs implementations or other ways to read and write on an Offline Feature Store.

The future work that could be carried out on the system evaluation could expand on one or more of these aspects: data, pipelines, and experimental environment. Expanding on data would mean running the experiments with larger tables (600M rows, 6BN rows, etc.) to explore and verify if a threshold where a Spark-based system performs better than delta-rs is present, and at which table size. Also making variations on the data sources would be a valid approach, although this would make this thesis an invalid baseline. The pipelines that were defined are specific to delta-rs or the Hopsworks architecture. Verifying the speed of other Offline Feature Stores as Databricks' would help generalize the results on a broader area. This would help clear out which approach performs best (between Spark and delta-rs) across different systems. The experimental environment as said in the limitations, was a shared environment with large resources, but varying on the current machine usage by other company's employees. Using clean isolated hardware could help future research verify this thesis's findings, isolating the variable results of a shared environment.

References

[1] "State of the Data Lakehouse," Dremio, Tech. Rep., 2024. [Page 1.]

[2] M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia, "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," in *Proceedings of CIDR*, vol. 8, 2021. [Pages 1 and 20.]

[3] D. Croci, "Data Lakehouse, beyond the hype," Dec. 2022. [Page 1.]

[4] "Apache Hudi vs Delta Lake vs Apache Iceberg - Data Lakehouse Feature Comparison," https://www.onehouse.ai/blog/apache-hudi-vs-delta-lake-vs-apache-iceberg-lakehouse-feature-comparison. [Page 1.]

[5] P. Rajaperumal, "Uber Engineering's Incremental Processing Framework on Hadoop," https://www.uber.com/blog/hoodie/, Mar. 2017. [Pages 1, 3, and 20.]

[6] "Apache Iceberg - Apache Iceberg™," https://iceberg.apache.org/. [Pages 1 and 20.]

[7] M. Armbrust, T. Das, L. Sun, B. Yavuz, S. Zhu, M. Murthy, J. Torres, H. Van Hovell, A. Ionescu, A. Łuszczak, M. Świtakowski, M. Szafrański, X. Li, T. Ueshin, M. Mokhtar, P. Boncz, A. Ghodsi, S. Paranjpye, P. Senster, R. Xin, and M. Zaharia, "Delta lake: High-performance ACID table storage over cloud object stores," *Proceedings of the VLDB Endowment*, vol. 13, no. 12, pp. 3411–3424, Aug. 2020. doi: 10.14778/3415478.3415560 [Pages 1, 3, 18, and 20.]

[8] 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, and I. Stoica, "Apache Spark: A unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, Oct. 2016. doi: 10.1145/2934664 [Pages 1, 3, and 23.]

[9] A. Khazanchi, "Faster reading with DuckDB and arrow flight on hopsworks : Benchmark and performance evaluation of offline feature stores," Master's thesis, KTH Royal Institute of Technology / KTH, School of Electrical Engineering and Computer Science (EECS) / KTH, School of Electrical Engineering and Computer Science (EECS), 2023. [Pages 1 and 5.]

[10] M. Raasveldt and H. Mühleisen, "DuckDB: An Embeddable Analytical Database," in *Proceedings of the 2019 International Conference on Management of Data*. Amsterdam Netherlands: ACM, Jun. 2019. doi: 10.1145/3299869.3320212. ISBN 978-1-4503-5643-5 pp. 1981–1984. [Pages 2, 4, 24, and 38.]

[11] R. Vink, "I wrote one of the fastest DataFrame libraries," https://www.ritchievink.com/blog/2021/02/28/i-wrote-one-of-the-fastest-dataframe-libraries/, Feb. 2021. [Pages 2 and 4.]

[12] "Benchmark Results for Spark, Dask, DuckDB, and Polars — TPC-H Benchmarks at Scale," https://tpch.coiled.io/. [Page 2.]

[13] T. Ebergen, "Updates to the H2O.ai db-benchmark!" https://duckdb.org/2023/11/03/db-benchmark-update.html, Nov. 2023. [Page 2.]

[14] A. Nagpal and G. Gabrani, "Python for Data Analytics, Scientific and Technical Applications," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*, Feb. 2019. doi: 10.1109/AICAI.2019.8701341 pp. 140–145. [Page 2.]

[15] "TIOBE Index," https://www.tiobe.com/tiobe-index/. [Pages 2 and 4.]

[16] "Stack Overflow Developer Survey 2023," https://survey.stackoverflow.co/2023/?utm_source=social-share&utm_medium=social&utm_campaign=dev-survey-2023. [Page 2.]

[17] M. Raschka and S. Vahid, *Python Machine Learning (3rd Edition)*. Packt Publishing, 2019. ISBN 978-1-78995-575-0 [Page 2.]

[18] "Hopsworks - Batch and Real-time ML Platform," https://www.hopsworks.ai/, 2024. [Pages 2 and 4.]

[19] A. Pettersson, "Resource-efficient and fast Point-in-Time joins for Apache Spark : Optimization of time travel operations for the creation of machine learning training datasets," Master's thesis, KTH, School of Electrical Engineering and Computer Science (EECS) / KTH, School of Electrical Engineering and Computer Science (EECS), 2022. [Page 2.]

[20] "What Is a Lakehouse?" https://www.databricks.com/blog/2020/01/30/what-is-a-data-lakehouse.html, Thu, 01/30/2020 - 09:00. [Page 3.]

[21] S. Chaudhuri and U. Dayal, "An overview of data warehousing and OLAP technology," *ACM SIGMOD Record*, vol. 26, no. 1, pp. 65–74, Mar. 1997. doi: 10.1145/248603.248616 [Page 3.]

[22] EDER, "Unstructured Data and the 80 Percent Rule," Aug. 2008. [Page 3.]

[23] "Dremel made simple with Parquet," https://blog.x.com/engineering/en_us/a/2013/dremel-made-simple-with-parquet. [Page 3.]

[24] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, ser. HotCloud'10. USA: USENIX Association, 2010, p. 10. [Pages 3 and 21.]

[25] "Apache Spark™ - Unified Engine for large-scale data analytics," https://spark.apache.org/. [Page 3.]

[26] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache Flink™: Stream and Batch Processing in a Single Engine." [Page 3.]

[27] G. van Rossum, "Python tutorial," Centrum voor Wiskunde en Informatica (CWI), Amsterdam, Tech. Rep. CS-R9526, May 1995. [Page 4.]

[28] "TIOBE Index," https://www.tiobe.com/tiobe-index/programminglanguages_definition/. [Page 4.]

[29] "Delta-io/delta-rs," Delta Lake, May 2024. [Pages 5, 6, 36, 45, and 48.]

[30] S. Niazi, M. Ismail, S. Haridi, J. Dowling, S. Grohsschmiedt, and M. Ronström, "HopsFS: Scaling Hierarchical File System Metadata Using NewSQL Databases," in *15th USENIX Conference on File and Storage Technologies (FAST 17)*, 2017. ISBN 978-1-931971-36-2 pp. 89–104. [Pages 5, 9, and 16.]

[31] "The Apache Spark Open Source Project on Open Hub," https://openhub.net/p/apache-spark. [Page 6.]

[32] "Green Software Foundation," https://greensoftware.foundation/. [Page 7.]

[33] "What is Green Software?" https://greensoftware.foundation/articles/what-is-green-software, Oct. 2021. [Page 7.]

[34] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon Emissions and Large Neural Network Training," 2021. [Page 7.]

[35] D. Patterson, J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink," 2022. [Page 7.]

[36] "| Sustainable Development," https://sdgs.un.org/. [Page 7.]

[37] M. Frampton, *Complete Guide to Open Source Big Data Stack*, Jan. 2018. ISBN 978-1-4842-2149-5 [Page 11.]

[38] S. Sakr, "Big Data Processing Stacks," *IT Professional*, vol. 19, no. 1, pp. 34–41, Jan. 2017. doi: 10.1109/MITP.2017.6 [Page 11.]

[39] "Block vs File vs Object Storage - Difference Between Data Storage Services - AWS," https://aws.amazon.com/compare/the-difference-between-block-file-object-storage/. [Page 12.]

[40] "How Object vs Block vs File Storage differ," https://cloud.google.com/discover/object-vs-block-vs-file-storage. [Page 12.]

[41] "Object vs. File vs. Block Storage: What's the Difference?" https://www.ibm.com/blog/object-vs-file-vs-block-storage/, Oct. 2021. [Page 12.]

[42] D. Borthakur, "The Hadoop Distributed File System: Architecture and Design," 2005. [Pages 15 and 23.]

[43] "Logicalclocks/rondb," Hopsworks, Oct. 2024. [Pages 16 and 26.]

[44] J. de la Rúa Martínez, F. Buso, A. Kouzoupis, A. A. Ormenisan, S. Niazi, D. Bzhalava, K. Mak, V. Jouffrey, M. Ronström, R. Cunningham, R. Zangis, D. Mukhedkar, A. Khazanchi, V. Vlassov, and J. Dowling, "The hopsworks feature store for machine learning," in *Companion of the 2024 International Conference on Management of Data*, ser. Sigmod/Pods '24. New York, NY, USA: Association for Computing Machinery, 2024. doi: 10.1145/3626246.3653389. ISBN 9798400704222 pp. 135–147. [Page 17.]

[45] H. E. Pence, "What is Big Data and Why is it Important?" *Journal of Educational Technology Systems*, vol. 43, no. 2, pp. 159–171, Dec. 2014. doi: 10.2190/ET.43.2.d [Page 18.]

[46] I. Gorton and J. Klein, "Distribution, Data, Deployment: Software Architecture Convergence in Big Data Systems," *IEEE Software*, vol. 32, no. 3, pp. 78–85, May 2015. doi: 10.1109/MS.2014.51 [Page 18.]

[47] "Parquet," https://parquet.apache.org/. [Page 21.]

[48] "Announcing Delta Lake 3.0 with New Universal Format and Liquid Clustering," https://www.databricks.com/blog/announcing-delta-lake-30-new-universal-format-and-liquid-clustering, Thu, 06/29/2023 - 06:53. [Page 21.]

[49] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, San Francisco, CA, 2004, pp. 137–150. [Page 23.]

[50] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing," EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2011-82, Jul. 2011. [Page 23.]

[51] J. Kreps, N. Narkhede, and J. Rao, "Kafka: A Distributed Messaging System for Log Processing," *Proceedings of the NetDB*, 2011. [Page 23.]

[52] R. Shah and R. Tkachuk, "Improve Your OLAP Environment with Microsoft and Teradata." [Page 24.]

[53] wesm, "Introducing Apache Arrow Flight: A Framework for Fast Data Transport," https://arrow.apache.org/blog/2019/10/13/introducing-arrow-flight/, Oct. 2019. [Page 25.]

[54] "gRPC," https://grpc.io/. [Page 25.]

[55] "Meet Michelangelo: Uber's Machine Learning Platform," https://www.uber.com/en-RO/blog/michelangelo-machine-learning-platform/, Sep. 2017. [Page 25.]

[56] M. L. Despa, "Comparative study on software development methodologies." *Database Systems Journal*, vol. 5, no. 3, 2014. [Page 32.]

[57] "Welcome home : Vim online," https://www.vim.org/. [Page 34.]

[58] "Neoclide/coc.nvim," Neoclide, Sep. 2024. [Page 34.]

[59] H. Fann, "Fannheyward/coc-rust-analyzer," Sep. 2024. [Page 34.]

[60] "Docker Build," https://docs.docker.com/build/, 12:14:28 +0200 +0200. [Page 35.]

[61] "GitHub," https://github.com. [Page 35.]

[62] "TPC-H Homepage," https://www.tpc.org/tpch/. [Page 38.]

[63] T. P. P. C. (TPC), "TPC-H_v3.0.1.pdf," 1993/2022. [Page 38.]

[64] M. Poess and C. Floyd, "New TPC benchmarks for decision support and web commerce," *Sigmod Record*, vol. 29, no. 4, pp. 64–71, Dec. 2000. doi: 10.1145/369275.369291 [Page 38.]

[65] A. Behm, S. Palkar, U. Agarwal, T. Armstrong, D. Cashman, A. Dave, T. Greenstein, S. Hovsepian, R. Johnson, A. Sai Krishnan, P. Leventis, A. Luszczak, P. Menon, M. Mokhtar, G. Pang, S. Paranjpye, G. Rahn, B. Samwel, T. Van Bussel, H. Van Hovell, M. Xue, R. Xin, and M. Zaharia, "Photon: A Fast Query Engine for Lakehouse Systems," in *Proceedings of the 2022 International Conference on Management of Data*. Philadelphia PA USA: ACM, Jun. 2022. doi: 10.1145/3514221.3526054. ISBN 978-1-4503-9249-5 pp. 2326–2339. [Page 38.]

[66] "TPC Current Specs," https://www.tpc.org/tpc_documents_current_versions/current_specific [Page 38.]

[67] "Arrow-rs/object_store/README.md at master · apache/arrow-rs," https://github.com/apache/arrow-rs/blob/master/object_store/README.md. [Page 46.]

[68] A. Binford, "Kimahriman/hdfs-native," Aug. 2024. [Pages 47 and 67.]

[69] "Rustls/tokio-rustls: Async TLS for the Tokio runtime," https://github.com/rustls/tokio-rustls. [Page 48.]

[70] G. Manfredi, "Silemo/hdfs-native," Aug. 2024. [Page 48.]

[71] ——, "Silemo/hdfs-native-object-store," Aug. 2024. [Page 48.]

[72] ——, "Silemo/delta-rs," Sep. 2024. [Page 48.]

[73] "Delta-rs/python at main · Silemo/delta-rs," https://github.com/Silemo/delta-rs/tree/main/python. [Page 48.]

[74] A. Balasubramanian, M. S. Baranowski, A. Burtsev, A. Panda, Z. Rakamarić, and L. Ryzhyk, "System programming in rust: Beyond safety," in *Proceedings of the 16th Workshop on Hot Topics in Operating Systems*, ser. HotOS '17. New York, NY, USA: Association for Computing Machinery, 2017. doi: 10.1145/3102980.3103006. ISBN 978-1-4503-5068-6 pp. 156–161. [Page 63.]

Appendix A

System architectures

This appendix reports the legacy architecture diagrams shown in Section 2.5 increased in size to improve readability.

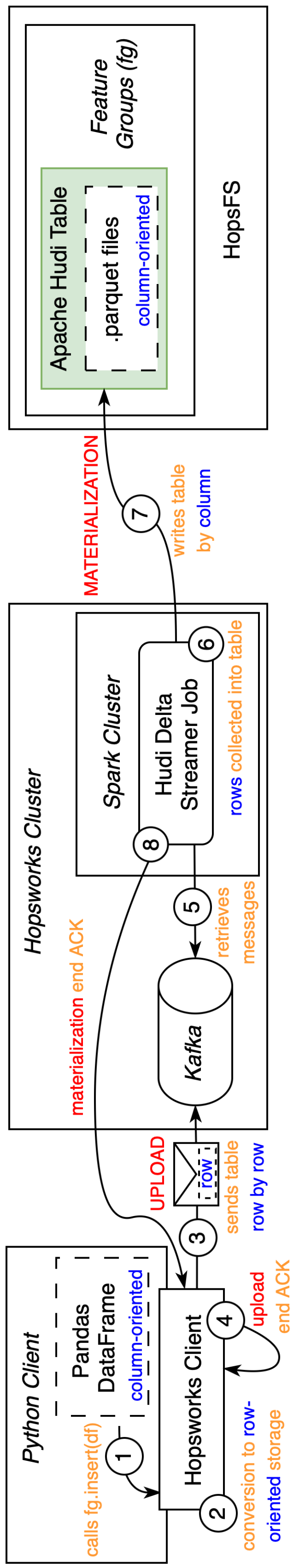

Figure A.1: Legacy system writing process

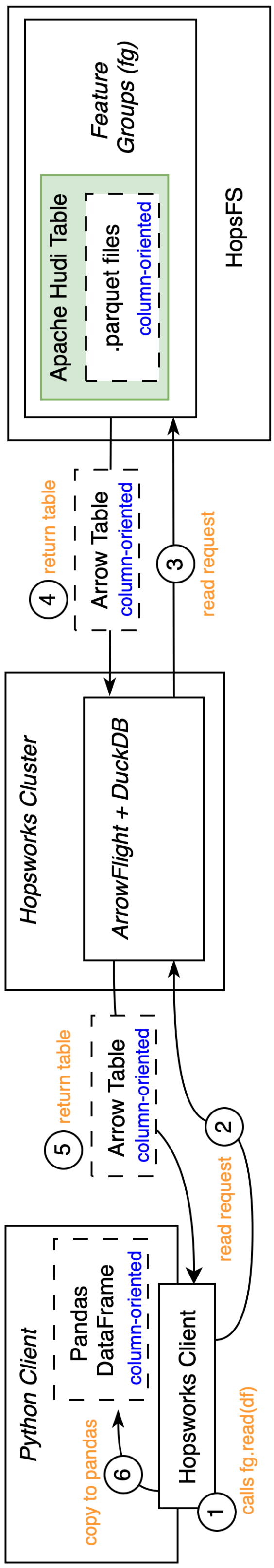

Figure A.2: Legacy system reading process

Appendix B

Write experiments results

This appendix reports all graphs and tables related to all the writing experiments conducted. Results are reported first expressed as latency (measured during the experiments), then as throughput (computed from the latency and table size).

Table B.1: Write latency experiment results performed with 1 CPU core

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
delta-rs HopsFS	10K	1.250 88	1.238 07	1.265 45
	100K	1.368 28	1.337 57	1.389 82
	1M	9.381 52	9.239 71	9.529 04
	6M	19.754 69	19.332 70	20.117 85
	60M	177.307 07	174.628 71	180.017 32
delta-rs LocalFS	10K	0.039 57	0.037 70	0.041 53
	100K	0.152 40	0.145 98	0.158 88
	1M	8.422 52	8.283 96	8.563 76
	6M	17.906 34	17.480 40	18.335 85
	60M	172.345 52	169.748 08	174.731 38
Legacy	10K	50.227 67	49.535 01	50.936 64
	100K	59.561 87	58.894 66	60.184 96
	1M	112.190 48	111.371 62	113.009 15
	6M	511.816 93	510.751 13	512.836 72
	60M	2 715.772 85	2 699.880 61	2 731.952 25

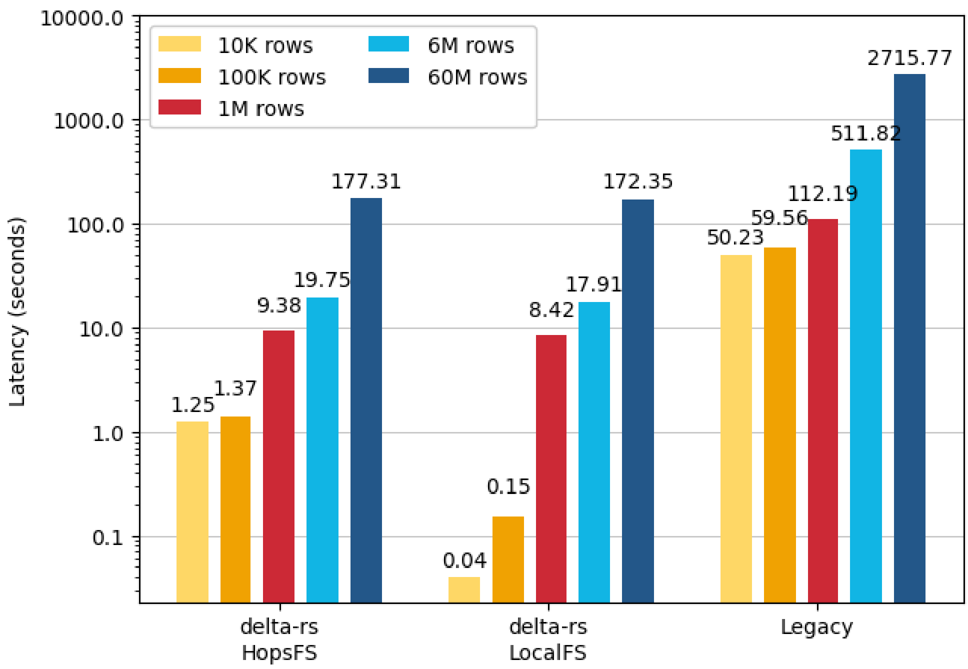

Figure B.1: Histogram in log-scale of the write latency experiment results performed with 1 CPU core

Table B.2: Write latency experiment results performed with 2 CPU cores

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval	
			low	high
delta-rs HopsFS	10K	1.262 39	1.250 79	1.276 39
	100K	1.308 12	1.280 50	1.332 17
	1M	8.515 36	8.343 33	8.700 77
	6M	16.290 42	15.906 59	16.673 62
	60M	134.060 89	131.650 31	136.397 61
delta-rs LocalFS	10K	0.048 23	0.046 40	0.049 97
	100K	0.137 14	0.134 02	0.140 50
	1M	7.185 30	7.037 47	7.351 28
	6M	15.266 32	14.851 72	15.651 67
	60M	129.820 07	127.600 20	132.046 89
Legacy	10K	50.724 05	50.107 69	51.306 86
	100K	59.788 10	58.979 97	60.474 27
	1M	108.564 99	108.011 24	109.081 28
	6M	473.379 54	472.345 34	474.437 40
	60M	2 340.770 13	2 333.994 43	2 347.971 27

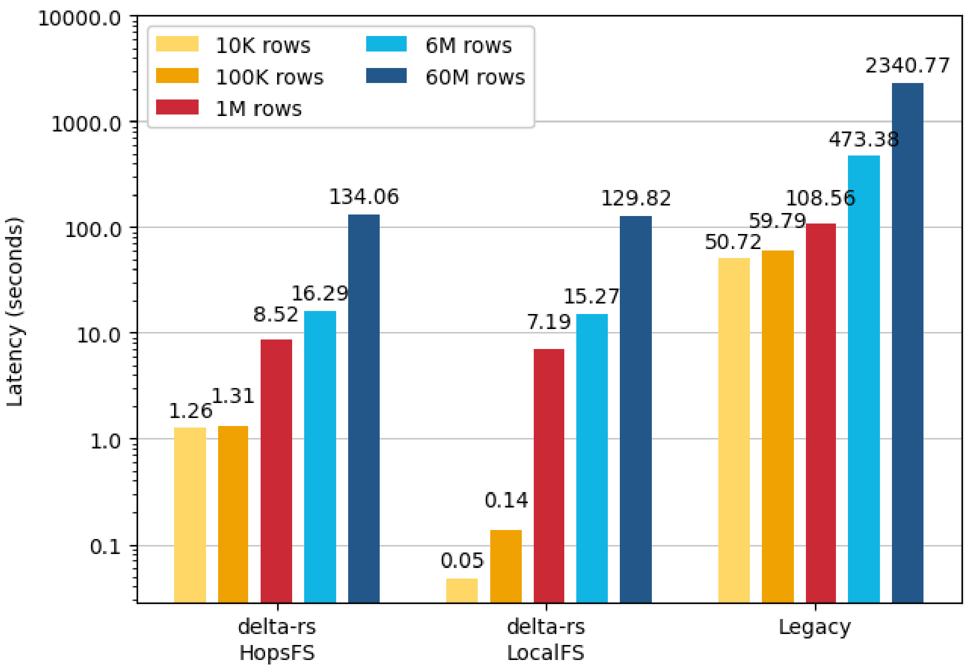

Figure B.2: Histogram in log-scale of the write latency experiment results performed with 2 CPU cores

Table B.3: Write latency experiment results performed with 4 CPU cores

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
delta-rs HopsFS	10K	1.216 42	1.202 32	1.232 31
	100K	1.336 22	1.322 94	1.349 42
	1M	8.413 25	8.247 70	8.582 72
	6M	16.224 02	15.879 46	16.595 86
	60M	124.102 42	121.577 23	126.815 30
delta-rs LocalFS	10K	0.045 72	0.043 41	0.048 07
	100K	0.131 76	0.128 80	0.134 99
	1M	7.185 74	7.006 79	7.363 43
	6M	14.555 78	14.176 79	14.941 92
	60M	121.376 23	119.172 56	123.698 90
Legacy	10K	51.284 65	50.622 82	51.903 67
	100K	59.526 55	58.905 37	60.153 22
	1M	108.816 74	108.252 17	109.342 34
	6M	481.983 53	481.044 35	482.929 92
	60M	2 346.046 87	2 336.993 96	2 355.198 97

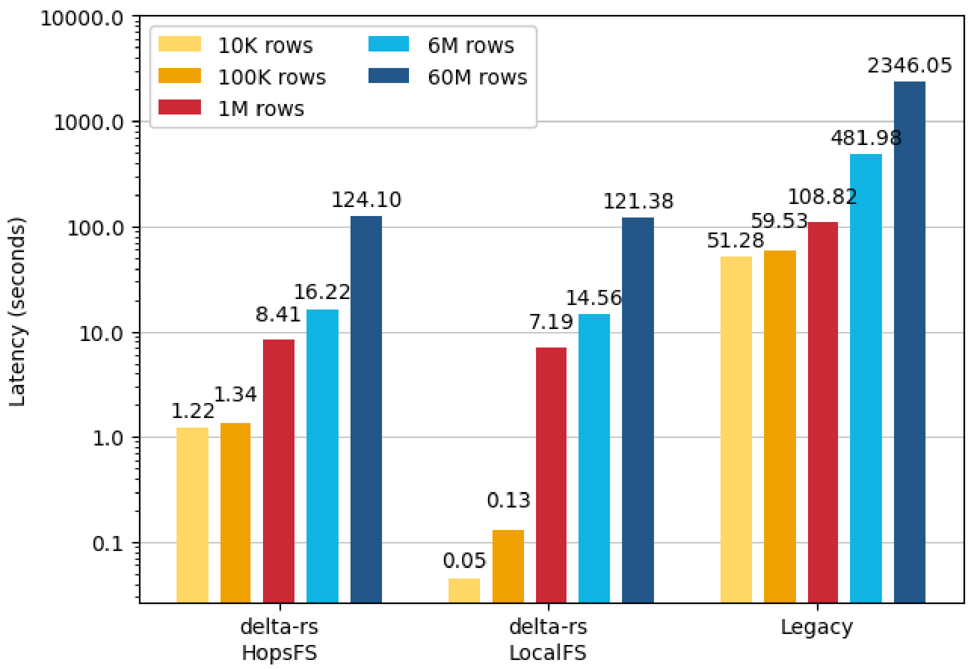

Figure B.3: Histogram in log-scale of the write latency experiment results performed with 4 CPU cores

Table B.4: Write latency experiment results performed with 8 CPU cores

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
delta-rs HopsFS	10K	1.367 56	1.249 34	1.572 24
	100K	1.292 43	1.265 48	1.310 99
	1M	8.301 20	8.149 18	8.470 40
	6M	15.738 47	15.289 74	16.160 84
	60M	121.950 14	119.593 76	124.180 97
delta-rs LocalFS	10K	0.044 02	0.041 74	0.046 40
	100K	0.136 48	0.132 81	0.140 61
	1M	7.228 72	7.075 11	7.398 93
	6M	14.281 57	13.905 08	14.661 26
	60M	119.979 15	117.764 16	122.208 82
Legacy	10K	51.228 59	50.594 78	51.864 76
	100K	60.277 51	59.729 07	60.771 30
	1M	109.381 89	108.868 30	109.882 63
	6M	475.943 45	474.839 93	477.052 74
	60M	2 324.979 17	2 319.042 03	2 331.047 94

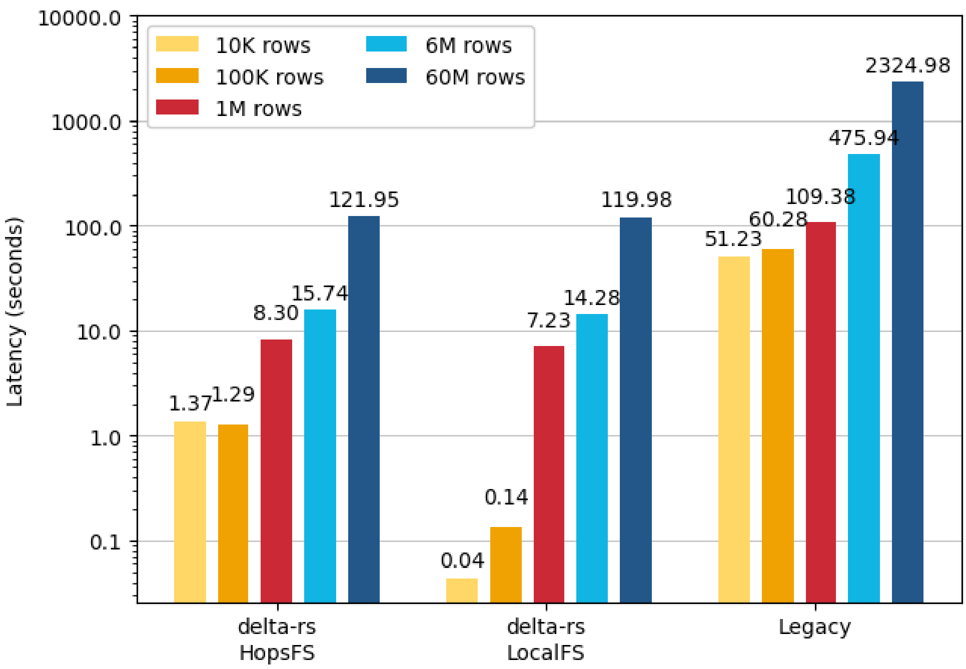

Figure B.4: Histogram in log-scale of the write latency experiment results performed with 8 CPU cores

Table B.5: Write throughput experiment results performed with 1 CPU core

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	7.994 36	7.902 30	8.077 05
	100K	7.308 43	7.195 14	7.476 21
	1M	106.592 42	104.942 26	108.228 50
	6M	303.725 33	298.242 52	310.354 91
	60M	338.395 98	333.301 26	343.586 10
delta-rs LocalFS	10K	252.682 38	240.734 05	265.186 32
	100K	656.157 39	629.367 77	684.985 18
	1M	118.729 19	116.771 10	120.715 14
	6M	335.076 75	327.227 70	343.241 43
	60M	348.137 84	343.384 22	353.464 96
Legacy	10K	0.199 09	0.196 32	0.201 87
	100K	1.678 92	1.661 54	1.697 94
	1M	8.913 41	8.848 84	8.978 94
	6M	11.722 94	11.699 63	11.747 40
	60M	22.093 15	21.962 31	22.223 20

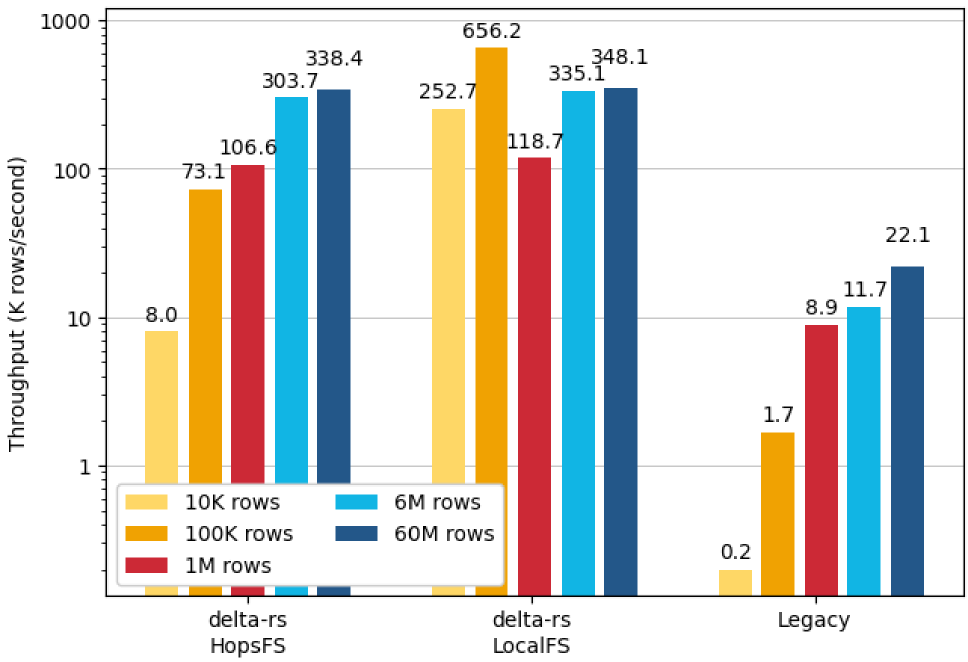

Figure B.5: Histogram in log-scale of the write throughput experiment results performed with 1 CPU core

Table B.6: Write throughput experiment results performed with 2 CPU cores

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	7.921 46	7.834 58	7.994 91
	100K	76.445 07	75.064 99	78.094 33
	1M	117.434 78	114.932 31	119.856 14
	6M	368.314 40	359.849 75	377.201 98
	60M	447.557 80	439.890 38	455.752 81
delta-rs LocalFS	10K	207.319 22	200.086 26	215.493 42
	100K	729.159 67	711.739 66	746.138 54
	1M	139.172 97	136.030 55	142.096 47
	6M	393.021 85	383.345 60	403.993 52
	60M	462.178 14	454.384 03	470.218 68
Legacy	10K	0.197 14	0.194 90	0.199 57
	100K	1.672 57	1.653 59	1.695 49
	1M	9.211 07	9.167 47	9.258 29
	6M	12.674 81	12.646 55	12.702 57
	60M	25.632 58	25.553 97	25.707 00

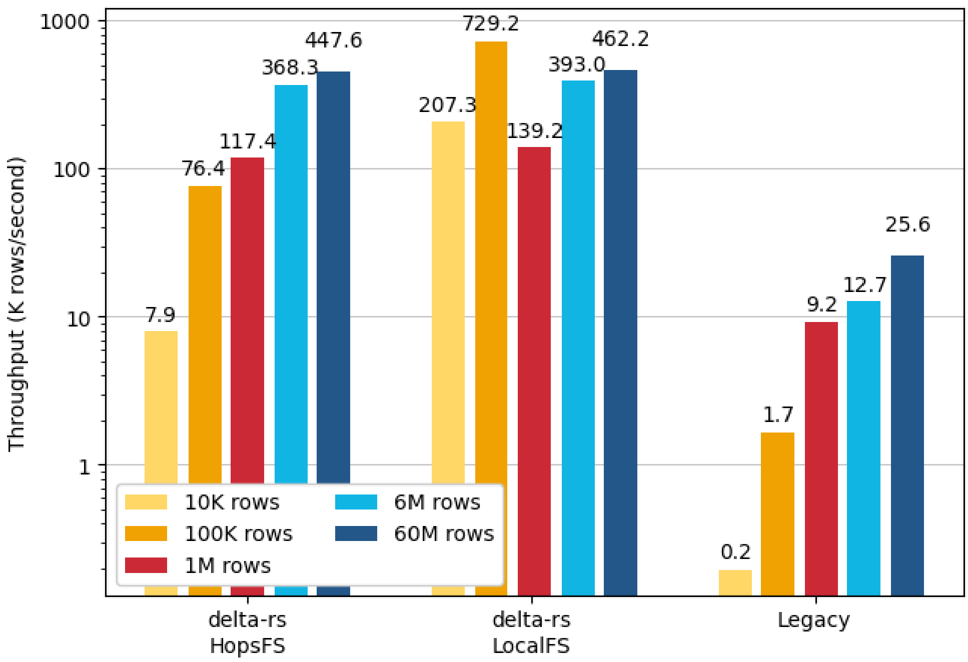

Figure B.6: Histogram in log-scale of the write throughput experiment results performed with 2 CPU cores

Table B.7: Write throughput experiment results performed with 4 CPU cores

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	8.220 83	8.114 82	8.114 82
	100K	74.837 42	74.105 66	75.588 71
	1M	118.860 08	116.513 10	121.245 82
	6M	369.822 02	361.535 77	377.846 52
	60M	483.471 60	473.128 99	493.513 43
delta-rs LocalFS	10K	218.713 64	208.026 04	230.314 12
	100K	758.924 22	740.794 74	776.391 15
	1M	139.164 32	135.806 13	142.718 64
	6M	412.207 28	401.554 59	423.226 78
	60M	494.330 70	485.048 75	503.471 56
Legacy	10K	0.194 99	0.192 66	0.197 53
	100K	1.679 92	1.662 42	1.697 63
	1M	9.189 76	9.145 58	9.237 68
	6M	12.448 55	12.424 16	12.472 86
	60M	25.574 93	25.475 55	25.674 00

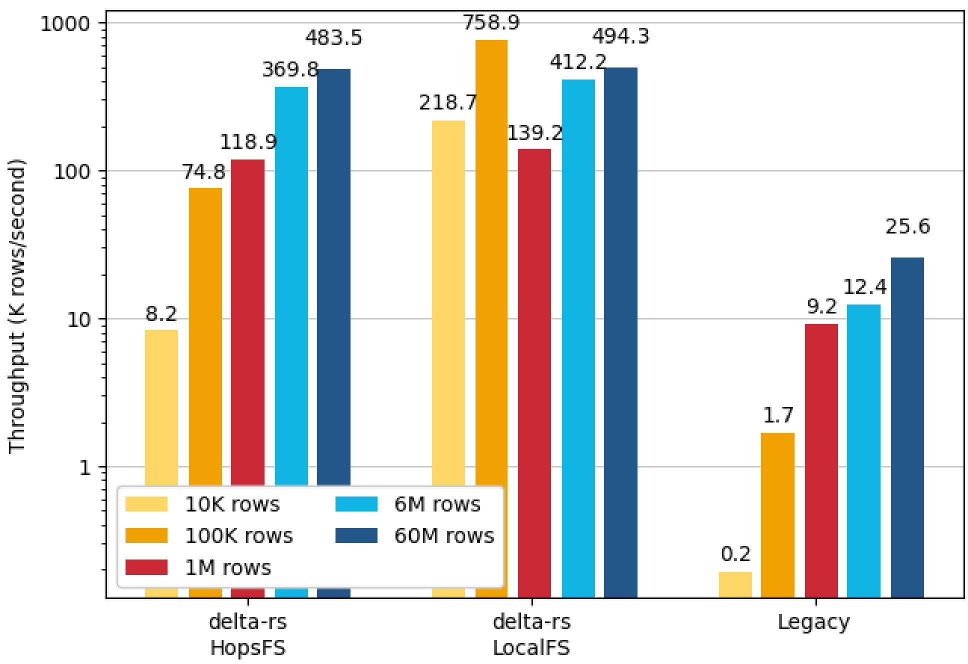

Figure B.7: Histogram in log-scale of the write throughput experiment results performed with 4 CPU cores

Table B.8: Write throughput experiment results performed with 8 CPU cores

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	7.312 28	6.360 32	8.004 22
	100K	77.373 37	76.277 82	79.021 04
	1M	120.464 39	118.058 14	122.711 60
	6M	381.231 26	371.267 72	392.419 78
	60M	492.004 31	483.165 79	501.698 37
delta-rs LocalFS	10K	227.120 95	215.487 14	239.541 28
	100K	732.701 41	711.160 38	752.932 00
	1M	138.337 01	135.154 66	141.340 41
	6M	420.121 65	409.241 76	431.496 69
	60M	500.086 88	490.962 88	509.492 86
Legacy	10K	0.195 20	0.192 80	0.197 64
	100K	1.658 99	1.645 51	1.674 22
	1M	9.142 28	9.100 61	9.185 40
	6M	12.606 53	12.577 22	12.635 83
	60M	25.806 68	25.739 49	25.872 75

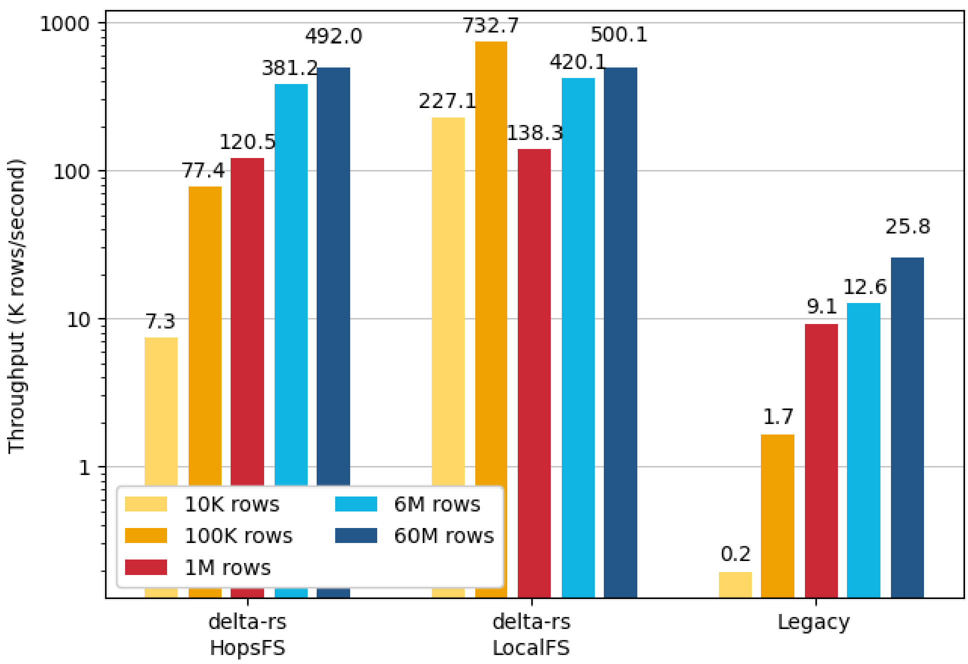

Figure B.8: Histogram in log-scale of the write throughput experiment results performed with 8 CPU cores

Appendix C

Read experiments results

This appendix reports all graphs and tables related to all the reading experiments conducted. Results are reported first expressed as latency (measured during the experiments), then as throughput (computed from the latency and table size).

Table C.1: Read latency experiment results performed with 1 CPU core

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval	
			low	high
delta-rs HopsFS	10K	0.053 42	0.039 16	0.081 12
	100K	0.057 57	0.055 18	0.060 46
	1M	0.538 55	0.525 58	0.552 29
	6M	1.948 99	1.930 07	1.968 60
	60M	22.980 65	22.840 67	23.142 06
delta-rs LocalFS	10K	0.004 19	0.002 68	0.006 44
	100K	0.026 96	0.019 66	0.034 33
	1M	0.420 09	0.406 13	0.435 63
	6M	1.682 23	1.659 81	1.704 40
	60M	19.565 47	19.346 90	19.777 24
Legacy	10K	0.631 59	0.624 14	0.641 57
	100K	2.650 10	2.642 72	2.658 76
	1M	8.596 36	8.340 94	8.900 47
	6M	33.529 64	33.238 86	33.865 91
	60M	33.697 72	33.362 62	34.086 65

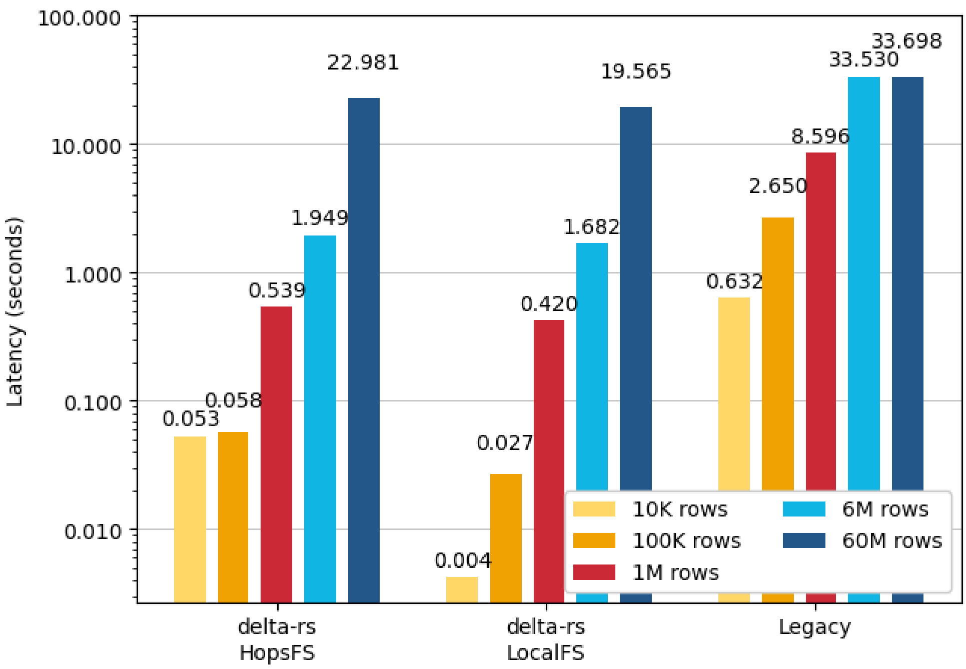

Figure C.1: Histogram in log-scale of the read latency experiment results performed with 1 CPU core

Table C.2: Read latency experiment results performed with 2 CPU cores

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
delta-rs HopsFS	10K	0.041 32	0.039 33	0.043 78
	100K	0.056 90	0.051 23	0.066 93
	1M	0.234 13	0.225 28	0.244 26
	6M	0.908 32	0.899 67	0.917 44
	60M	11.413 25	11.276 61	11.588 99
delta-rs LocalFS	10K	0.002 87	0.002 78	0.002 99
	100K	0.013 06	0.010 41	0.016 10
	1M	0.199 77	0.188 58	0.210 56
	6M	0.747 64	0.735 03	0.760 13
	60M	9.446 93	9.372 07	9.517 53
Legacy	10K	0.624 92	0.622 10	0.628 22
	100K	2.663 39	2.656 16	2.671 66
	1M	8.616 67	8.309 89	8.949 38
	6M	33.375 19	33.096 88	33.670 65
	60M	33.642 81	33.301 50	34.063 07

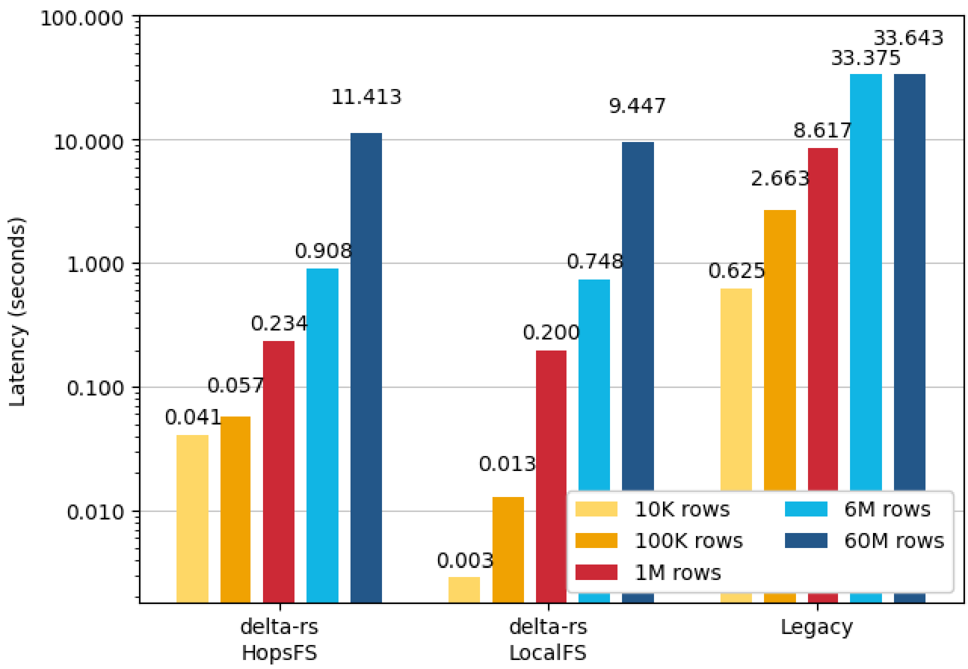

Figure C.2: Histogram in log-scale of the read latency experiment results performed with 2 CPU cores

Table C.3: Read latency experiment results performed with 4 CPU cores

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
delta-rs HopsFS	10K	0.043 36	0.039 22	0.050 92
	100K	0.055 40	0.053 78	0.057 89
	1M	0.188 47	0.151 57	0.251 89
	6M	0.531 24	0.507 78	0.571 05
	60M	5.580 11	5.543 97	5.619 36
delta-rs LocalFS	10K	0.002 68	0.002 59	0.002 79
	100K	0.009 23	0.008 52	0.010 20
	1M	0.089 71	0.083 88	0.095 03
	6M	0.370 21	0.360 18	0.380 32
	60M	4.810 23	4.793 38	4.827 89
Legacy	10K	0.635 83	0.623 52	0.659 08
	100K	2.639 85	2.633 49	2.646 23
	1M	8.752 38	8.507 25	9.013 83
	6M	33.452 86	33.194 61	33.756 46
	60M	33.652 45	33.270 16	34.039 00

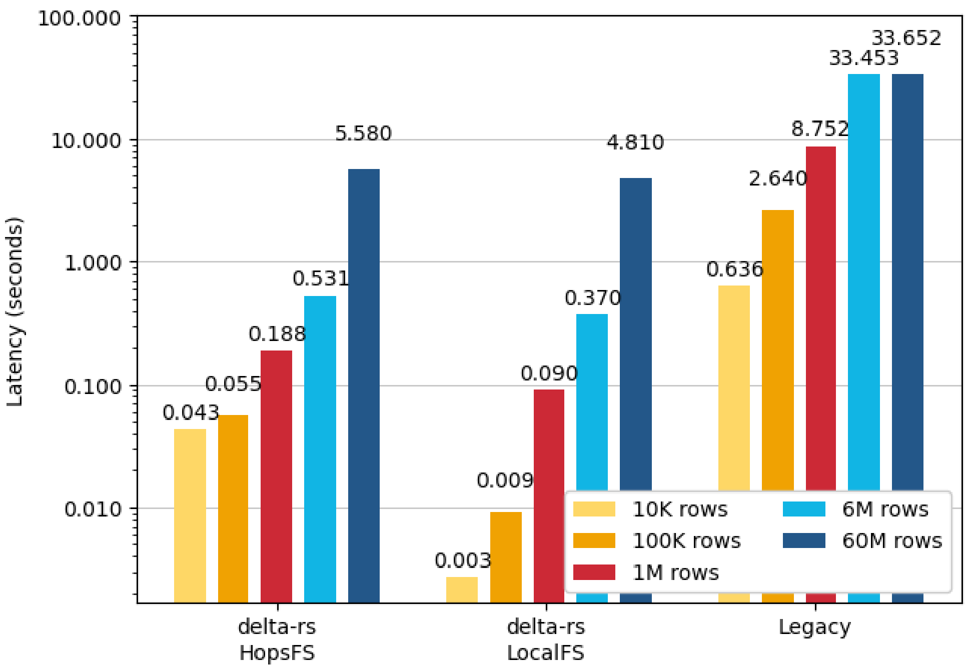

Figure C.3: Histogram in log-scale of the read latency experiment results performed with 4 CPU cores

Table C.4: Read latency experiment results performed with 8 CPU cores

Pipeline	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
delta-rs HopsFS	10K	0.043 30	0.038 85	0.051 19
	100K	0.054 58	0.052 86	0.056 63
	1M	0.173 90	0.169 92	0.178 08
	6M	0.497 29	0.486 55	0.510 19
	60M	2.942 36	2.855 54	3.063 60
delta-rs LocalFS	10K	0.002 94	0.002 84	0.003 07
	100K	0.009 48	0.008 72	0.010 56
	1M	0.043 08	0.039 34	0.048 30
	6M	0.175 48	0.170 09	0.180 80
	60M	2.285 50	2.275 01	2.296 07
Legacy	10K	0.627 39	0.622 45	0.632 59
	100K	2.662 17	2.653 09	2.672 18
	1M	8.347 57	8.134 71	8.609 42
	6M	33.428 15	33.153 76	33.749 47
	60M	33.143 41	32.883 03	33.412 99

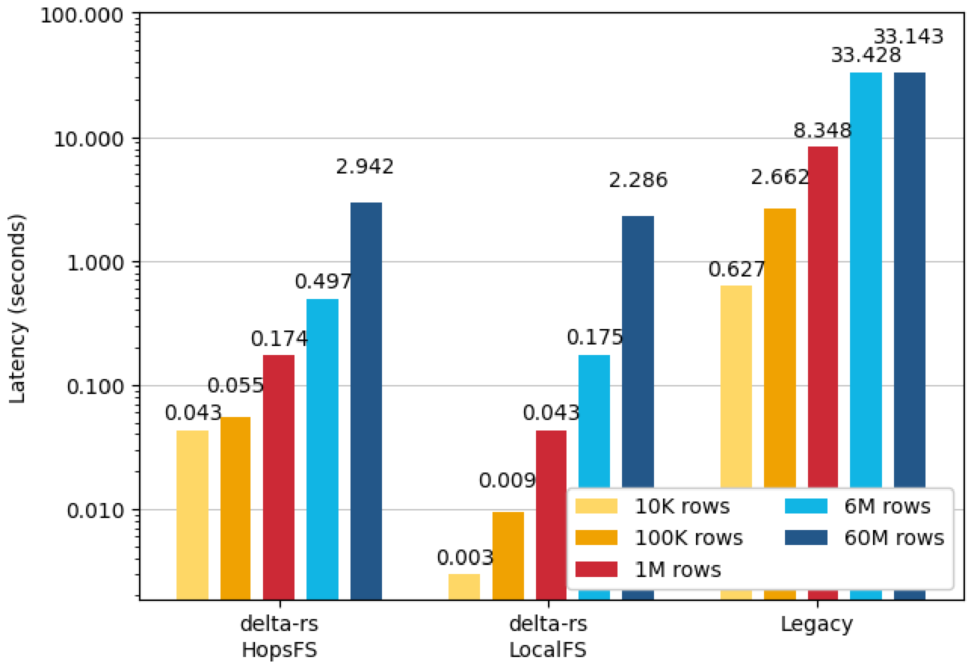

Figure C.4: Histogram in log-scale of the read latency experiment results performed with 8 CPU cores

Table C.5: Read throughput experiment results performed with 1 CPU core

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	187.168 53	123.265 55	255.361 73
	100K	1 736.907 99	1 653.919 40	1 811.928 57
	1M	1 856.831 67	1 810.613 18	1 902.653 61
	6M	3 078.512 99	3 047.839 14	3 108.694 31
	60M	2 610.891 46	2 592.681 85	2 626.892 40
delta-rs LocalFS	10K	2 384.586 99	1 552.085 52	3 721.360 68
	100K	3 708.257 87	2 912.436 00	5 084.757 15
	1M	2 380.403 81	2 295.501 54	2 462.258 14
	6M	3 566.674 54	3 520.297 96	3 614.870 06
	60M	3 066.626 44	3 033.789 96	3 101.271 65
Legacy	10K	15.832 85	15.586 54	16.021 96
	100K	37.734 32	37.611 40	37.839 75
	1M	116.328 20	112.353 50	119.890 44
	6M	178.946 12	177.169 28	180.511 59
	60M	1 780.535 63	1 760.219 74	1 798.419 62

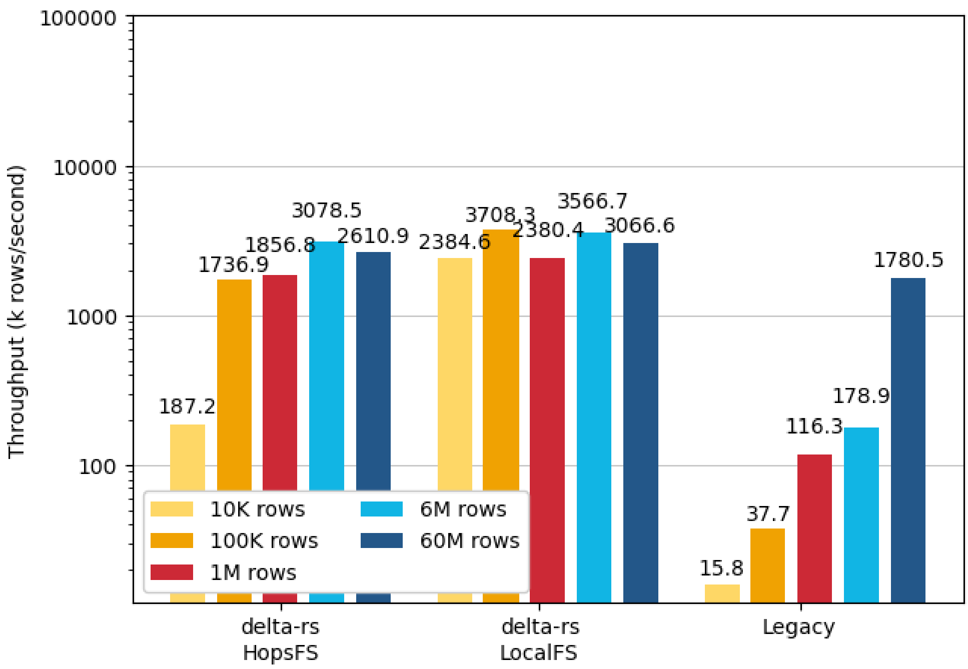

Figure C.5: Histogram in log-scale of the read throughput experiment results performed with 1 CPU core

Table C.6: Read throughput experiment results performed with 2 CPU cores

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	241.978 33	228.370 18	254.224 19
	100K	1 757.179 30	1 494.018 76	1 951.812 29
	1M	4 271.121 39	4 093.989 17	4 438.846 12
	6M	6 605.543 23	6 539.936 31	6 669.087 56
	60M	5 257.046 58	5 177.323 95	5 320.743 34
delta-rs LocalFS	10K	3 479.952 85	3 339.119 82	3 592.422 55
	100K	7 652.747 09	6 210.296 38	9 604.332 93
	1M	5 005.616 42	4 749.179 43	5 302.536 56
	6M	8 025.196 34	7 893.332 51	8 162.844 93
	60M	6 351.267 15	6 304.155 38	6 401.994 67
Legacy	10K	16.001 83	15.917 73	16.074 53
	100K	37.546 07	37.429 79	37.648 22
	1M	116.054 04	111.739 51	120.338 48
	6M	179.774 24	178.196 71	181.285 93
	60M	1 783.441 98	1 761.438 30	1 801.720 13

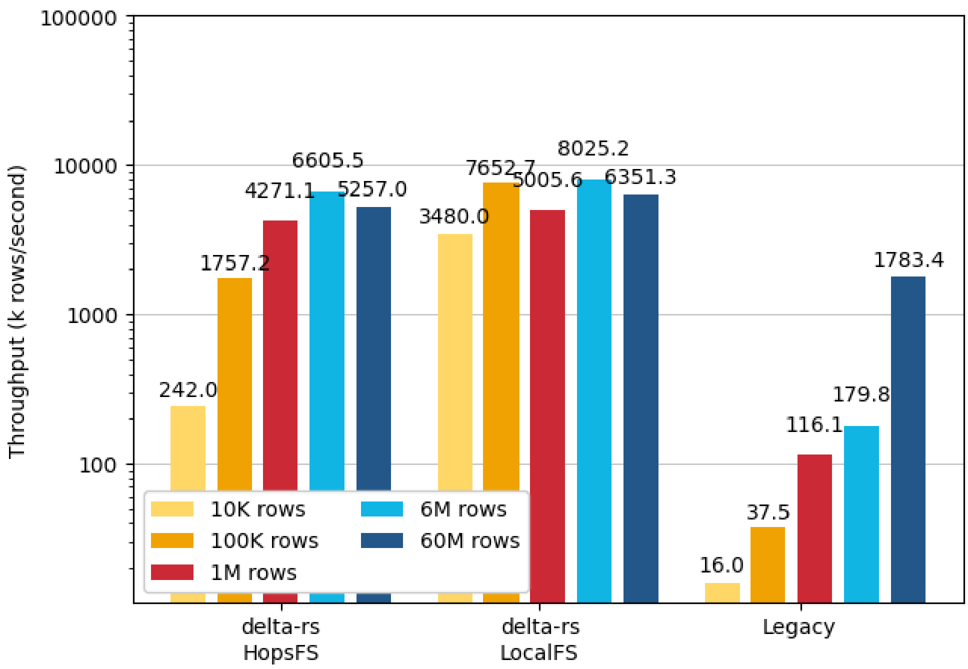

Figure C.6: Histogram in log-scale of the read throughput experiment results performed with 2 CPU cores

Table C.7: Read throughput experiment results performed with 4 CPU cores

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	230.619 25	196.373 66	254.932 32
	100K	1 804.731 06	1 727.272 56	1 859.407 19
	1M	5 305.742 39	3 969.956 99	6 597.374 04
	6M	11 294.126 19	10 506.919 39	11 816.033 13
	60M	10 752.475 11	10 677.356 47	10 822.556 20
delta-rs LocalFS	10K	3 720.941 04	3 572.450 85	3 854.749 73
	100K	10 830.884 57	9 802.960 62	11 735.928 63
	1M	11 146.067 43	10 522.540 71	11 921.622 57
	6M	16 206.973 30	15 775.943 87	16 657.937 25
	60M	12 473.404 92	12 427.787 28	12 517.259 67
Legacy	10K	15.727 24	15.172 59	16.037 90
	100K	37.880 93	37.789 59	37.972 37
	1M	114.254 60	110.940 61	117.546 71
	6M	179.356 80	177.743 72	180.752 22
	60M	1 782.930 73	1 762.683 96	1 803.417 64

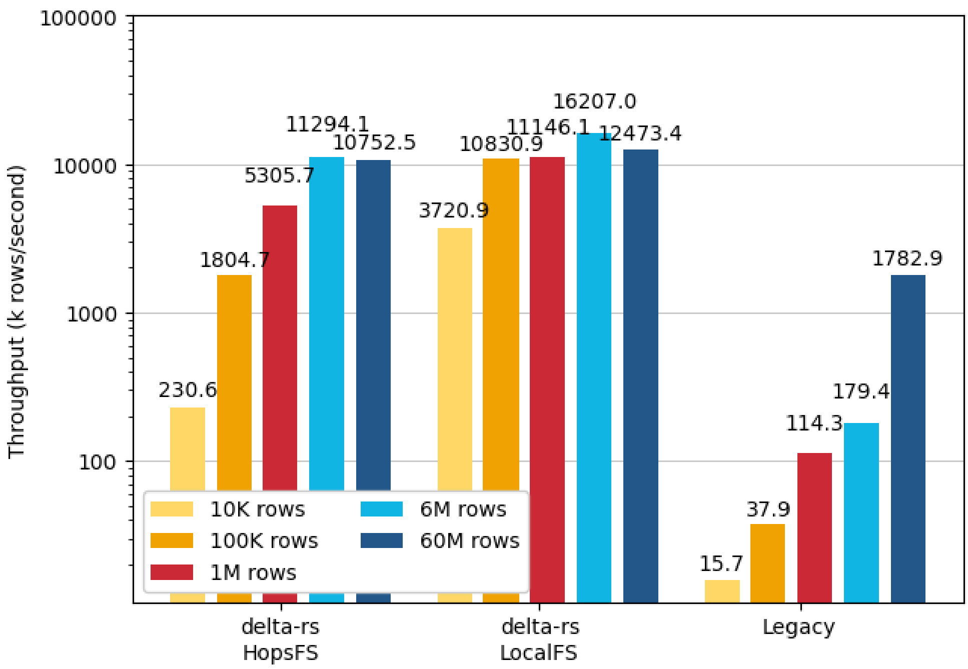

Figure C.7: Histogram in log-scale of the read throughput experiment results performed with 4 CPU cores

Table C.8: Read throughput experiment results performed with 8 CPU cores

Pipeline	Number of rows	Throughput (k rows/second)	Throughput (k rows/second) 95% Confidence Interval low	high
delta-rs HopsFS	10K	230.925 18	195.345 44	257.360 67
	100K	1 832.056 83	1 765.756 84	1 891.544 56
	1M	5 750.343 66	5 615.170 71	5 885.017 95
	6M	12 065.182 02	11 760.178 93	12 331.709 77
	60M	20 391.779 56	19 584.753 63	21 011.721 36
delta-rs LocalFS	10K	3 390.322 42	3 256.074 86	3 510.692 31
	100K	10 545.410 87	9 465.275 36	11 463.060 73
	1M	23 212.466 79	20 701.230 33	25 414.131 22
	6M	34 191.396 37	33 185.181 29	35 275.254 56
	60M	26 252.420 19	26 131.518 36	26 373.488 80
Legacy	10K	15.939 01	15.807 91	16.065 39
	100K	37.563 30	37.422 50	37.691 86
	1M	119.795 20	116.151 77	122.929 95
	6M	179.489 42	177.780 54	180.974 90
	60M	1 810.314 36	1 795.708 67	1 824.648 83

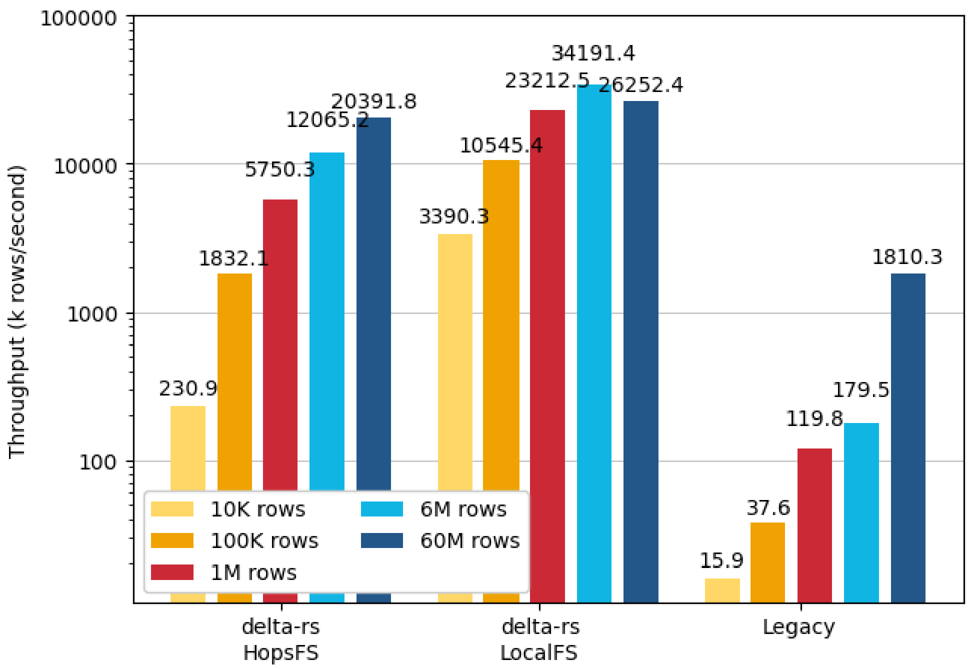

Figure C.8: Histogram in log-scale of the read throughput experiment results performed with 8 CPU cores

Appendix D

Legacy pipeline write latency breakdown results

This appendix reports all graphs and tables related to all write latency breakdown of the upload and materialization steps in the legacy pipeline.

Table D.1: Contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 1 CPU core

Step	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval	
			low	high
upload	10K	2.486 5	2.389 6	2.626 1
materialize		47.726 2	47.044 5	48.403 1
upload	100K	3.668 4	3.631 0	3.709 8
materialize		55.900 5	55.249 4	56.554 1
upload	1M	22.593 4	22.449 6	22.734 9
materialize		89.575 4	88.828 6	90.304 9
upload	6M	244.612 3	244.036 8	245.190 5
materialize		267.249 0	266.428 7	268.154 9
upload	6M	2 437.784 0	2 422.870 4	2 453.674 6
materialize		278.050 4	276.234 0	280.092 1

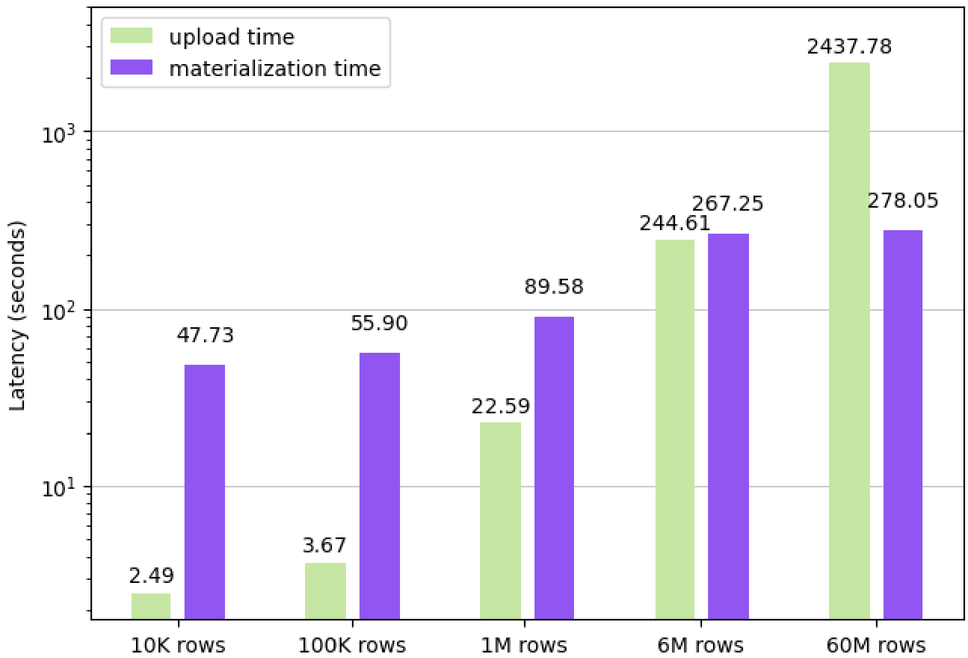

Figure D.1: Histogram in log-scale displaying the contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 1 CPU core

Table D.2: Contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 2 CPU cores

Step	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
upload	10K	2.387 3	2.327 6	2.446 6
materialize		48.330 5	47.702 0	48.992 3
upload	100K	3.434 8	3.400 8	3.467 1
materialize		56.336 7	55.562 6	57.112 9
upload	1M	18.634 9	18.567 3	18.710 4
materialize		89.926 7	89.401 2	90.451 4
upload	6M	205.885 4	205.217 7	206.498 4
materialize		267.507 9	266.685 3	268.351 2
upload	6M	2 064.135 7	2 057.639 6	2 070.445 0
materialize		276.960 8	275.715 6	278.284 9

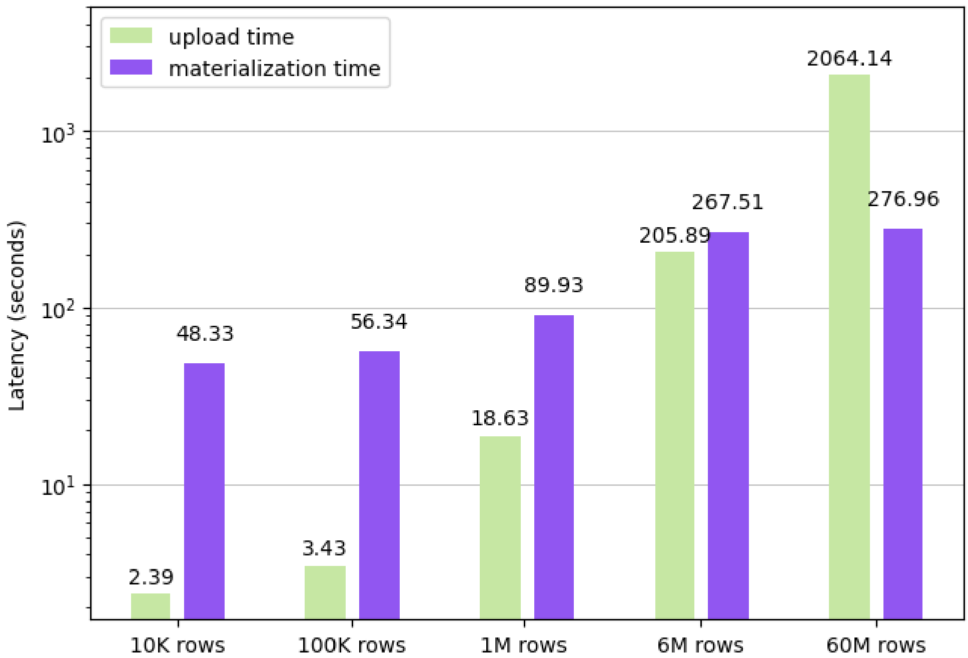

Figure D.2: Histogram in log-scale displaying the contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 2 CPU cores

Table D.3: Contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 4 CPU cores

Step	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
upload	10K	2.384 6	2.329 9	2.433 5
materialize		48.906 1	48.243 6	49.547 0
upload	100K	3.465 0	3.424 5	3.507 1
materialize		56.052 4	55.368 2	56.682 2
upload	1M	19.229 6	19.145 5	19.316 1
materialize		89.586 4	89.020 9	90.131 3
upload	6M	211.675 8	211.069 4	212.283 9
materialize		270.323 3	269.596 7	270.989 5
upload	6M	2 068.526 0	2 060.335 8	2 077.283 7
materialize		277.600 1	276.006 5	279.145 6

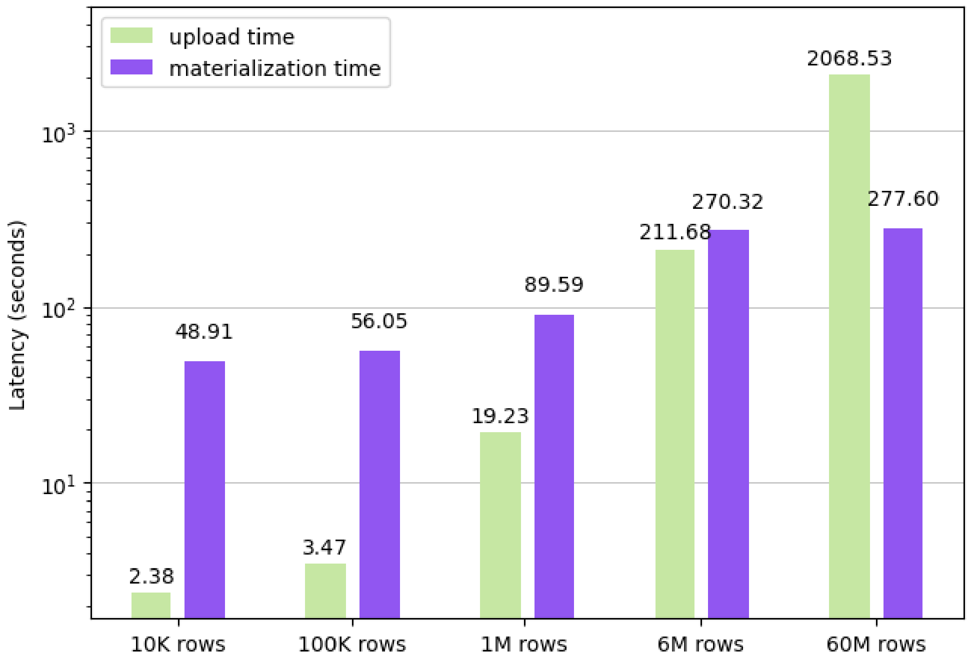

Figure D.3: Histogram in log-scale displaying the contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 4 CPU cores

Table D.4: Contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 8 CPU cores

Step	Number of rows	Latency (seconds)	Latency (seconds) 95% Confidence Interval low	high
upload materialize	10K	2.381 5 48.848 5	2.330 4 48.197 9	2.435 8 49.446 7
upload materialize	100K	3.439 2 56.842 8	3.408 1 56.317 7	3.470 0 57.368 5
upload materialize	1M	18.864 2 90.515 3	18.780 8 90.030 6	18.953 2 90.971 8
upload materialize	6M	207.664 6 268.275 2	207.160 6 267.356 9	208.209 0 269.245 6
upload materialize	6M	2 049.137 1 275.763 6	2 043.599 1 274.277 3	2 055.478 2 274.277 3

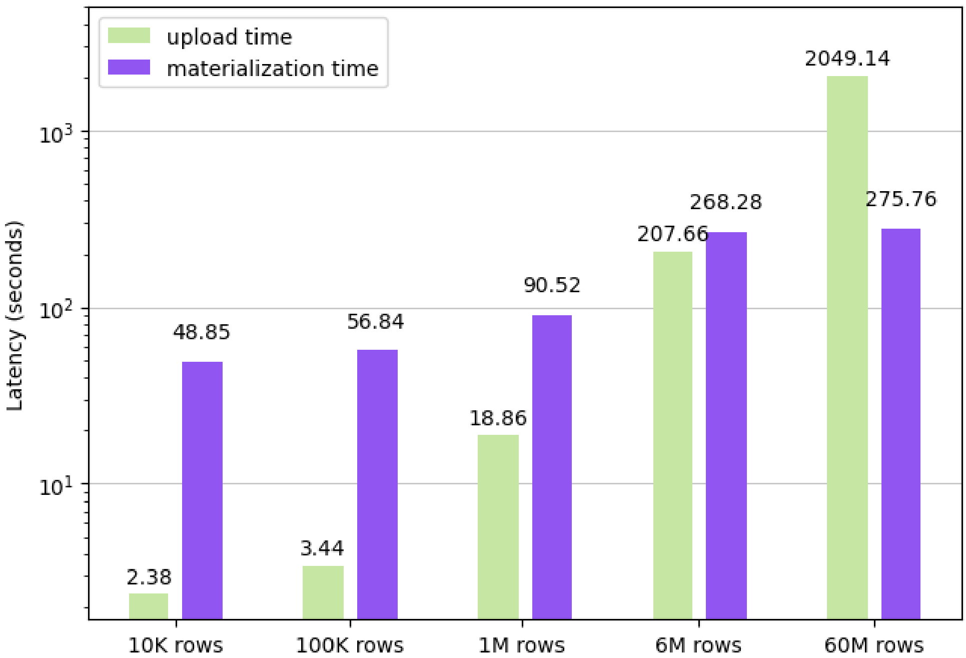

Figure D.4: Histogram in log-scale displaying the contributions to the write latency of the upload and materialization steps in the legacy pipeline performed with 8 CPU cores

TRITA – EECS-EX 2024:0000
Stockholm, Sweden 2024

www.kth.se

€€€€ For DIVA €€€€

{
"Author1": { "Last name": "Manfredi",
"First name": "Giovanni",
"Local User Id": "u142pmki",
"E-mail": "gioman@kth.se",
"organisation": {"L1": "School of Electrical Engineering and Computer Science",
}
},
"Cycle": "2",
"Course code": "DA258X",
"Credits": "30.0",
"Degree1": {"Educational program": "Master's Programme, ICT Innovation, 120 credits"
,"programcode": "TIVNM"
,"Degree": "Master's Programme, Distributed Systems and Data Mining for Big Data, 120 credits"
,"subjectArea": "?"
},
"Title": {
"Main title": "Reducing read and write latency using Rust in an offline feature store",
"Subtitle": "Developing delta-rs to support HopsFS and reducing read and write latency on the Hopsworks offline feature store",
"Language": "eng" },
"Alternative title": {
"Main title": "Detta är den svenska översättningen av titeln",
"Subtitle": "Detta är den svenska översättningen av undertiteln",
"Language": "swe"
},
"Supervisor1": { "Last name": "Sheikholeslami",
"First name": "Sina",
"Local User Id": "u1znylhw",
"E-mail": "sinash@kth.se",
"organisation": {"L1": "School of Electrical Engineering and Computer Science",
"L2": "Computer Science" }
},
"Supervisor2": { "Last name": "Schmidt",
"First name": "Fabian",
"Local User Id": "u1mrsz0u",
"E-mail": "fschm@kth.se",
"organisation": {"L1": "School of Electrical Engineering and Computer Science",
"L2": "Computer Science" }
},
"Supervisor3": { "Last name": "Niazi",
"First name": "Salman",
"E-mail": "salman@hopsworks.ai",
"Other organisation": "Hopsworks AB"
},
"Examiner1": { "Last name": "Vlassov",
"First name": "Vladimir",
"Local User Id": "u19yb2c8",
"E-mail": "vladv@kth.se",
"organisation": {"L1": "School of Electrical Engineering and Computer Science",
"L2": "Computer Science" }
},
"Cooperation": { "Partner_name": "Hopsworks AB"},
"National Subject Categories": "10201, 10206",
"Other information": {"Year": "2024", "Number of pages": "1,105"},
"Copyrightleft": "copyright",
"Series": { "Title of series": "TRITA – EECS-EX" , "No. in series": "2024:0000" },
"Opponents": { "Name": "A. B. Normal & A. X. E. Capion"},
"Presentation": { "Date": "2022-03-15 13:00"
,"Language":"eng"
,"Room": "via Zoom https://kth-se.zoom.us/j/ddddddddddd"
,"Address": "Isafjordsgatan 22 (Kistagången 16)"
,"City": "Stockholm" },
"Number of lang instances": "3",
"Abstract[eng]": €€€€
The need of building Machine Learning (ML) models based on ever increasing amounts of data brought new challenges to data management systems. Feature Stores have emerged as an effective solution to enable feature reuse, while organizing data transformations and ensuring consistency between feature engineering, model training and model inference. Recent publications demonstrate that the Hopsworks Feature Store exhibits superior performance metrics in both training and online inference query workloads compared to existing cloud-based alternatives. In this system, the latency to perform a write operation in this system a least of one or more minutes even for small quantities of data (1 GB or less). This limitation is believed to a Spark-specific limitation, which the system uses to write data on the offline feature store. This hypothesis was already confirmed in the case of read latency, where opting for a Spark-alternative, namely an Arrow Flight and DuckDB server, improved performance considerably. A promising approach appears to be adopting a new data management solution, namely Delta Lake, and accessing it using a Rust library called delta-rs. This thesis investigates the possibility of reducing the read and write latency in the offline feature store by expanding the delta-rs library to support the Hopsworks Feature Store file system called HopsFS, and comparatively evaluating the performance of the legacy and newly implemented system. After a first iterative system implementation phase based on fixed requirements, the system was evaluated by performing and measuring read and write operations in a four different CPU configurations (increasing the number of CPU cores up to eight).

Experiments were performed fifty times to estimate a confidence interval allowing a comparative evaluation of the systems. Results confirmed the superior performance of the delta-rs library over the Spark system in all write operations with a tenfold reduction in the latency. Delta-rs also surpassed the Spark-alternative in read operations with a tenfold reduction in the latency in all but the experiment with the largest table (60M rows), where the improvement is of a smaller factor. These findings encourage future research investigating Spark-alternative when optimizing performance in small scale (1 GB - 100 GB) data management systems. €€€€,
"Keywords[eng]": €€€€
Canvas Learning Management System, Docker containers, Performance tuning €€€€,
"Abstract[swe]": €€€€
Här ska jag skriva ett abstract som är på ca 250 och 350 ord (1/2 A4-sida) med följande komponenter:

- Vad är ämnesområdet? (valfritt) Presenterar ämnesområdet för projektet.
- Kort problemformulering
- Varför var detta problem värt en kandidat-/masteruppsats? (*i.e.*, varför är problemet både betydande och av en lämplig svårighetsgrad för ett kandidat-/masteruppsats-projekt? Varför har ingen annan löst det än?)
- Hur löste du problemet? Vad var din metod/insikt?
- Resultat/slutsatser/konsekvenser/påverkan: Vilka är dina viktigaste resultat/ slutsatser? Vad kommer andra att göra baserat på dina resultat? Vad kan göras nu när du är klar - som inte kunde göras innan ditt examensarbete var klart?

€€€€,
"Keywords[swe]": €€€€
Första nyckelordet, Andra nyckelordet, Tredje nyckelordet, Fjärde nyckelordet €€€€,
"Abstract[ita]": €€€€
Qui scriverò un abstract di circa 250 e 350 parole (1/2 pagina A4) con i seguenti elementi:

- Qual è l'area tematica? (opzionale) Introduce l'area tematica del progetto.
- Breve esposizione del problema
- Perché questo problema meritava un progetto di tesi di laurea/master? (Perché il problema è significativo e di un grado di difficoltà adeguato per un progetto di tesi di laurea/master? Perché nessun altro l'ha ancora risolto?)
- Come avete risolto il problema? Qual è stato il vostro metodo/intuizione?
- Risultati/Conclusioni/Conseguenze/Impatto: Quali sono i vostri risultati chiave/ conclusioni? Cosa faranno gli altri sulla base dei vostri risultati? Cosa si può fare ora che avete finito - che non si poteva fare prima che il vostro progetto di tesi fosse completato?

€€€€,
"Keywords[ita]": €€€€
Prima parola chiave, Seconda parola chiave, Terza parola chiave, Quarta parola chiave €€€€,
}

acronyms.tex

```
%%% Local Variables:
%%% mode: latex
%%% TeX-master: t
%%% End:
%% The following command is used with glossaries-extra
\setabbreviationstyle[acronym]{long-short}
%% The form of the entries in this file is \newacronym{label}{acronym}{phrase}
%%                                      or \newacronym[options]{label}{acronym}{phrase}
%% see "User Manual for glossaries.sty" for the  details about the options, one example is shown below
%% note the specification of the long form plural in the line below
%%\newacronym[longplural={Debugging Information Entities}]{DIE}{DIE}{Debugging Information Entity}
%%
%% The following example also uses options
%%\newacronym[shortplural={OSes}, firstplural={operating systems (OSes)}]{OS}{OS}{operating system}
%%
%% note the use of a non-breaking dash in long text for the following acronym
%%\newacronym{IQL}{IQL}{Independent Q^^e2^^80^^91Learning}
%%
%% Notes
%% 1. you can't use the \gls() command in a heading - but you can get the short (\glsentryshort)
%% or long version (\glsentryshort) or \glsentrylong or even the text entry (\glsentrytext) and then there is

\newacronym{KTH}{KTH}{KTH Royal Institute of Technology}
\newacronym{AB}{AB}{\textit{Aktiebolag}, tr. Limited company}
\newacronym{ACID}{ACID}{Atomicity, Consistency, Isolation and Durability}
\newacronym{AI}{AI}{Artificial Intelligence}
\newacronym{ML}{ML}{Machine Learning}
\newacronym{BI}{BI}{Business Intelligence}
\newacronym[shortplural={RDDs}, firstplural={Resilient Distributed Datasets (RDDs)}]{RDD}{RDD}{Resilient Dist
\newacronym{OLAP}{OLAP}{On-Line Analytical Processing}
\newacronym{ELT}{ELT}{Extract Load Transform}
\newacronym{ETL}{ETL}{Extract Transform Load}
\newacronym{HDFS}{HDFS}{Hadoop Distributed File System}
\newacronym{JVM}{JVM}{Java Virtual Machine}
\newacronym[shortplural={INs}, firstplural={Industrial Needs (INs)}]{IN}{IN}{Industrial Need}
\newacronym[shortplural={PAs}, firstplural={Project Assumptions (PAs)}]{PA}{PA}{Project Assumption}
\newacronym[shortplural={APIs}, firstplural={Application Programming Interfaces (APIs)}]{API}{API}{Applicatio
\newacronym{OLTP}{OLTP}{On-Line Transaction Processing}
\newacronym[shortplural={DBMs}, firstplural={Data Base Management Systems}]{DBMS}{DBMS}{Data Base Management
\newacronym[shortplural={Gs}, firstplural={Goals}]{G}{G}{Goal}
\newacronym[shortplural={RQs}, firstplural={Research Questions}]{RQ}{RQ}{Research Question}
\newacronym[shortplural={Ds}, firstplural={Deliverables}]{D}{D}{Deliverable}
\newacronym{CRUD}{CRUD}{Create Read Update Delete}
\newacronym[shortplural={SDGs}, firstplural={Sustainable Development Goals}]{SDG}{SDG}{Sustainable Developmen
\newacronym{AWS}{AWS}{Amazon Web Services}
\newacronym{GCS}{GCS}{Google Cloud Storage}
\newacronym[shortplural={HDDs}, firstplural={Hard Disks Drives}]{HDD}{HDD}{Hard Disk Drive}
\newacronym[shortplural={SSDs}, firstplural={Solid State Drives}]{SSD}{SSD}{Solid State Drive}
\newacronym[shortplural={PCs}, firstplural={Personal Computers}]{PC}{PC}{Personal Computer}
\newacronym[shortplural={OSes}, firstplural={Operating Systems (OSes)}]{OS}{OS}{Operating System}
\newacronym[shortplural={DFSes}, firstplural={Distributed File Systems (DFSes)}]{DFS}{DFS}{Distributed File S
\newacronym{BPMN}{BPMN}{Business Process Model and Notation}
\newacronym{TLS}{TLS}{Trasport Layer Security}
\newacronym{SSH}{SSH}{Secure Shell protocol}
\newacronym{VM}{VM}{Virtual Machine}
\newacronym{CoC}{CoC}{Conquer of Completion}
\newacronym{SF}{SF}{Scale Factor}
\newacronym{CPU}{CPU}{Central Processing Unit}
\newacronym{GPU}{GPU}{Graphical Processing Unit}
\newacronym{HopsFS}{HopsFS}{Hopsworks' \glsentryshort{HDFS} distribution}
\newacronym{LocalFS}{LocalFS}{Local File System}
\newacronym{TPC}{TPC}{Transaction Processing Performance Council}
\newacronym{RAM}{RAM}{Random Access Memory}
\newacronym{RPC}{RPC}{Remote Procedural Call}
\newacronym{CIDR}{CIDR}{Conference on Innovative Data Systems Research}
\newacronym{MLOps}{MLOps}{Machine Learning Operations}
```