Degree Project in ?

Second cycle, 30 credits

Faster Delta Lake operations using Rust

How Delta-rs beats Spark in a small scale Feature Store

GIOVANNI MANFREDI

Faster Delta Lake operations using Rust

How Delta-rs beats Spark in a small scale Feature Store

GIOVANNI MANFREDI

Master's Programme, ICT Innovation, 120 credits
Date: October 10, 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

Here I will write an abstract that is about 250 and 350 words (1/2 A4-page) with the following components:

- What is the topic area? (optional) Introduces the subject area for the project.
- Short problem statement
- Why was this problem worth a Bachelor's/Master's thesis project? (*i.e.*, why is the problem both significant and of a suitable degree of difficulty for a Bachelor's/Master's thesis project? Why has no one else solved it yet?)
- How did you solve the problem? What was your method/insight?
- Results/Conclusions/Consequences/Impact: What are your key results/ conclusions? What will others do based on your results? What can be done now that you have finished - that could not be done before your thesis project was completed?

Keywords

Canvas Learning Management System, Docker containers, Performance tuning First keyword, Second keyword, Third keyword, Fourth keyword

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

Canvas Lärplattform, Dockerbehållare, Prestandajustering 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

1 Introduction **1**
1.1 Background . . . 2
1.2 Problem . . . 5
1.2.1 Research Question . . . 5
1.3 Purpose . . . 6
1.4 Goals . . . 6
1.5 Ethics and Sustainability . . . 7
1.6 Research Methodology . . . 8
1.6.1 Delimitations . . . 9
1.7 Thesis Structure . . . 10

2 Background **11**
2.1 Data Storage . . . 12
2.1.1 File storage vs. Object storage vs. Block storage . . . 12
2.1.2 Hadoop Distributed File System . . . 15
2.1.3 HopsFS as an HDFS evolution . . . 17
2.1.4 HDFS alternatives: Cloud object stores . . . 17
2.2 Data Management . . . 17
2.3 Query engine . . . 17
2.4 Application - Hopsworks Feature Store . . . 17
2.5 Programming Languages . . . 17

3 Method **19**
3.1 System implementation – RQ1 . . . 19
3.1.1 Research Paradigm . . . 19
3.1.2 Development process . . . 20
3.1.3 Requirements . . . 21
3.1.4 Development environment . . . 22
3.2 System evaluation – RQ2 . . . 23

3.2.1 Research Paradigm . . . 23
3.2.2 Evaluation Process . . . 23
3.2.3 Industrial use case . . . 24
3.2.4 Dataset . . . 26
3.2.5 Experimental Design . . . 28
3.2.6 Experimental environment . . . 29
3.2.7 Evaluation Framework . . . 30
3.2.8 Assessing Reliability and Validity . . . 31

4 Implementation **33**
4.1 Software design and development . . . 33
4.1.1 First approach . . . 34
4.1.2 Final solution . . . 35
4.2 Software deployment and usage . . . 36
4.3 Experiments set-up . . . 37

5 Results and Analysis **39**
5.1 Major Results . . . 39
5.1.1 Writing Experiments . . . 40
5.1.2 Reading Experiments . . . 43
5.1.3 Legacy Spark pipeline write latency breakdown . . . 46
5.1.4 In-memory resources usage . . . 48
5.2 Results Analysis and Discussion . . . 48
5.2.1 delta-rs performance on HopsFS are similar to the library's performance on the LocalFS . . . 49
5.2.2 delta-rs outperforms the Legacy Spark pipeline in every experiment . . . 49
5.2.3 delta-rs scales well with increased CPU cores . . . 49
5.2.4 The upload time in the legacy Spark pipeline becomes the bottleneck as table size increases . . . 50

References **51**

List of Figures

1.1 Hadoop MapReduce and Apache Spark execution differences . 4
1.2 SDG supported by this thesis 8

2.1 Data stack abstraction for this project 11
2.2 HDFS architecture during read and write operations 16

3.1 BPMN diagram of the System implementation process 21
3.2 BPMN diagram of the System evaluation process 25

4.1 delta-rs library (v. 0.15.x) before and after adding HDFS support 33
4.2 Architecture of the first implementation approach 35
4.3 Architecture of the final implementation approach 36

5.1 Histogram (a) and Table (b) reporting the write latency experiment results also across different CPU configurations . . 41
5.2 Histogram (a) and Table (b) reporting the write throughput experiment results also across different CPU configurations . . 42
5.3 Histogram (a) and Table (b) reporting the read latency experiment results also across different CPU configurations . . 44
5.4 Histogram (a) and Table (b) reporting the read throughput experiment results also across different CPU configurations . . 45
5.5 Histogram (a) and Table (b) reporting the contributions to the write latency of the upload and materialization steps in the legacy Spark pipeline and it change at different CPU configurations 47

List of Tables

2.1 Data storage features comparison 15

Listings

3.1 Output of a *lscpu* bash command on the test environment. . . . 29
4.1 Writing a data frame on a Delta Table with delta-rs on Hadoop Distributed File System (HDFS) or Hopsworks' HDFS distribution (HopsFS) 37
4.2 Reading a data frame on a Delta Table with delta-rs on HDFS or HopsFS 37
4.3 Timeit usage to measure the time required to write a Delta Lake table to HopsFS. 38
4.4 A simple time difference approach to measure the time required to write a Delta Lake table to HopsFS. 38

List of acronyms and abbreviations

ACID	Atomicity, Consistency, Isolation and Durability
AI	Artificial Intelligence
API	Application Programming Interface
BI	Business Intelligence
BPMN	Business Process Model and Notation
CoC	Conquer of Completion
CPU	Central Processing Unit
D	Deliverable
DBMS	Data Base Management System
DFS	Distributed File System
ELT	Extract Load Transform
ETL	Extract Transform Load
G	Goal
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
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

Lakehouse systems are increasingly becoming the primary choice for running analytics in large-sized companies (that have more than 1000 employees) [1].

This recent architecture design called Lakehouse [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, 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.

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

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

While large communities back all three projects, Delta Lake is acknowledged as the de-facto Lakehouse solution [4]. This is mainly thanks to Databricks, which first promoted this new architecture over data lakes among their clients around 2020 [5].

As a data query and processing engine, Delta Lake is typically used with Apache Spark [6]. This approach is effective when processing large quantities of data (1 TB or more) over the cloud, but whether this approach is effective on small quantities of data (100 GB or less) remains to be investigated [7].

DuckDB [8], a Data Base Management System (DBMS) and Polars [9], 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 brings an increase in costs and computation time [10, 11].

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 [12]. Python is currently also the most popular general purpose programming language [13, 14] and it is by far the most used language for Machine Learning (ML) and Artificial Intelligence (AI) applications [15], 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 [14]. In this scenario, creating a Python client for Delta Lake would be beneficial as it 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 interface for Delta Lake directly benefits Hopsworks AB, the host company of this master thesis. Hopsworks AB develops a Feature Store for ML, a centralized, collaborative data platform that enables the storage and access of reusable features [16]. This architecture also supports point-in-time correct datasets from historical feature data [17].

This project here presented, aims to 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 pipeline is Apache Spark-based and the key hypothesis of the project is that a faster non-Apache Spark alternative is possible. If effective, Hopsworks will consider integrating this system implementation into the Hopsworks Feature Store (open source version), greatly improving the experience of Python users working on small quantities of data (between 10 GB and 100 GB). More generally, this work will outline the possibility of Apache Spark alternatives in small-scale (between 10 GB and 100 GB) use cases.

1.1 Background

A clear understanding of the background of this project comes from appreciating three different key aspects: Lakehouse development, Apache

Spark relevance and flows, and Python as an emergent language.

Lakehouse is a term coined by Databricks in 2020 [18], 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 [19], 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 [20]. 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. Lakehouse systems solved the problems of Data lakes by implementing data management and performance features on top of open data formats such as Parquet [21]. 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 [22] and then Delta Lake in 2020 [5].

Spark is a distributed computing framework used to support large-scale data-intensive applications [23]. 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 [24]. The paradigm was later implemented as an open-source project by Yahoo! engineers under the name of Hadoop MapReduce [25]. Spark significantly improved the performance of Hadoop MapReduce (10 times better in its first iteration) [23] thanks to its use of Resilient Distributed Datasets (RDDs) [26]. RDDs is a distributed memory abstraction that enables a lazy in-memory computation that is tracked through the use of lineage graphs, ultimately increasing fault tolerance [26]. This means that Spark avoids going back and forth between storage disks to store the computation results, as represented in Figure 1.1.

Spark, which is open-sourced under the Apache foundation as Apache Spark [27] (from now on simply Spark), has seen widespread success and adoption in various applications, becoming the de-facto data-intensive

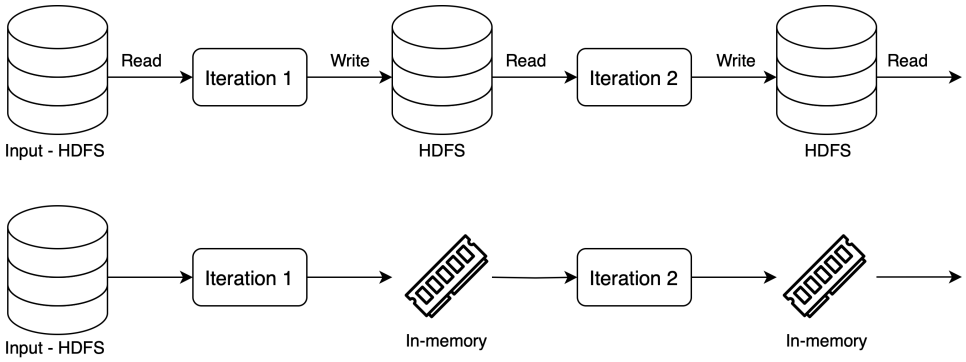


Figure 1.1: Hadoop MapReduce and Apache Spark execution differences

computing platform for the distributed computing world. While Spark is often used as a comprehensive solution [6], different solutions might be better suited for a specific scenario. An example of this is the case of Apache Flink [28], 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 the high-scaling capabilities of Spark may not be required. This is the case of DuckDB [8] and Polars [9], 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 [29]. 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 [30]. This is computed yearly in the TIOBE Index [13]. Looking at the 2024 list, it can be noted that Python has a rating of 15.16%, followed by C which has a rating of 10.97%. The index

also shows the trends of the last years, clearly displaying the rise of Python over historically very popular languages such as C and JAVA, which were both outranked by Python between 2021 and 2022. This shows the importance of offering Python APIs for programmers and data scientists in particular to increase the engagement and possibilities of a framework.

1.2 Problem

The Hopsworks Feature Store [16] when querying the Offline Feature Store, uses Spark as a query engine, i.e. executes the query (read, write, or delete) on the Offline Feature Store.

Currently, a write operation on the legacy Spark pipeline used in the Hopsworks Feature store will take one or more minutes to complete, even if writing a small quantity of data (only 1 GB of data or less). 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).

The research hypothesis that guides this research is that this slow transaction time is a Spark-specific issue. This is why Hopsworks has already adopted Spark alternatives [7] and it is experimenting with more Spark alternatives. Currently, Hopsworks is saving their Feature Store data on Apache Hudi and Delta Lake table formats. Delta Lake supports Spark alternatives for accessing and querying the data, and of particular interest is the delta-rs library [31] that enables Python access to Delta Lake tables, without having to use Spark. However, the delta-rs [31] does not support Hadoop Distributed File System (HDFS), and consequently Hopsworks' HDFS distribution (HopsFS) [32].

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 [31] operates on Delta Lake, using HopsFS [32]. To achieve this, support for HDFS (and thus also HopsFS) must be added to the delta-rs library [31], 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 (measured in seconds) and throughput (measured in rows/second) between a Spark-based operation (read or write) to Apache Hudi compared to a delta-rs library-based operation (read or write) to Delta Lake, in HopsFS?

1.3 Purpose

The purpose of this thesis project is 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 work will identify which one between a Spark pipeline and a delta-rs pipeline performs better at a small scale, by comparing the 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 will consider integrating this pipeline into their application.

Overall implications for this thesis work are much wider considering the popularity Spark has in the open source community (more than 2800 contributors during its lifetime [33]). This would enable developers to have a wider range of alternatives when working on "small scale" (1 GB to 100 GB) systems by choosing delta-rs over Spark.

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 [31] architecture and dependencies.

   G2: Identify what needs to be implemented to add HDFS support to the delta-rs library [31].

   G3: Implement HDFS support in the delta-rs library [31].

   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 Spark-based access to Apache Hudi compared a the delta-rs [31] 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 Spark-based access 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 [34] can be "part of the climate problem or part of the climate solution" [35]. 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 [35]. 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 [36, 37].

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 Spark-based 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) [38], 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 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 Spark pipeline has a

Figure 1.2: SDG supported by this thesis

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:

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.

These assumptions will be validated in Chapter 2.

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 [32]. Then, an evaluation structure will be designed and used to compare the performances of the old Spark-based 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

Once the thesis is written, provide an outline of the thesis structure

Chapter 2

Background

This project works on a layered data stack, that handles Big Data, i.e. large volumes of various 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 [39, 40], 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.

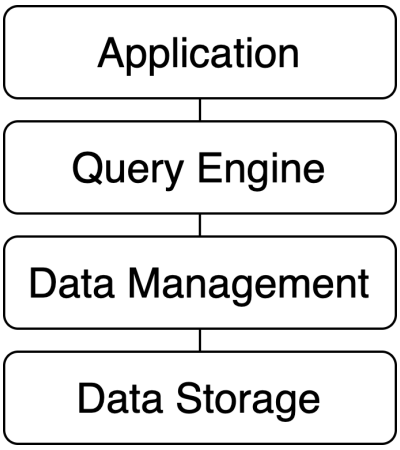


Figure 2.1: Data stack abstraction for this project

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/or support structured and/or 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/or 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 programming languages complements the Background by explaining the roles different programming languages play in the data stack (namely, Python, Java, Scala, and Rust).

2.1 Data Storage

This section aims to describe 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) [41, 42, 43].

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) [25].

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.

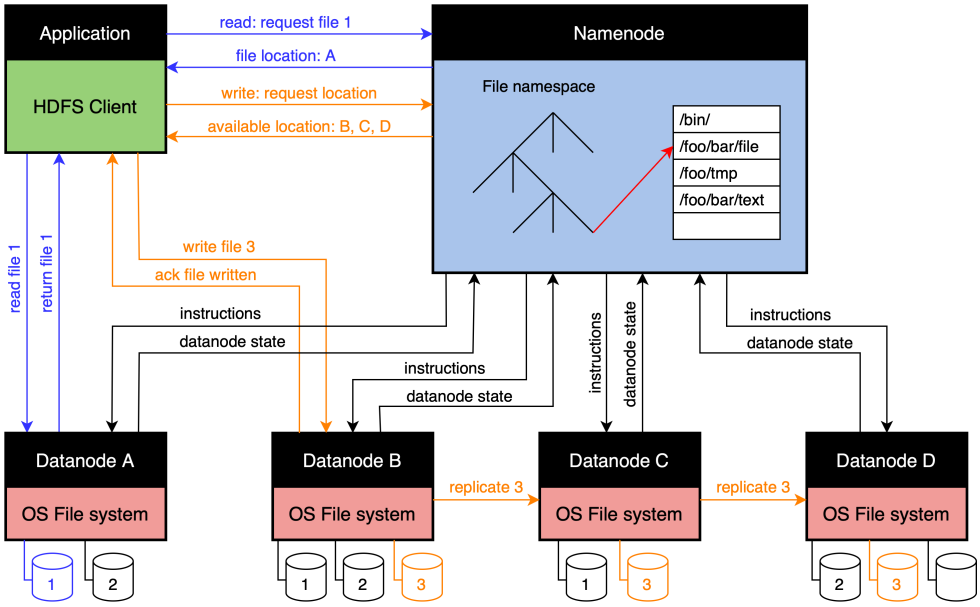


Figure 2.2: HDFS architecture during read and write operations

2.1.3 HopsFS as an HDFS evolution

2.1.4 HDFS alternatives: Cloud object stores

2.2 Data Management

2.3 Query engine

2.4 Application - Hopsworks Feature Store

2.5 Programming Languages

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 [44]. 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.

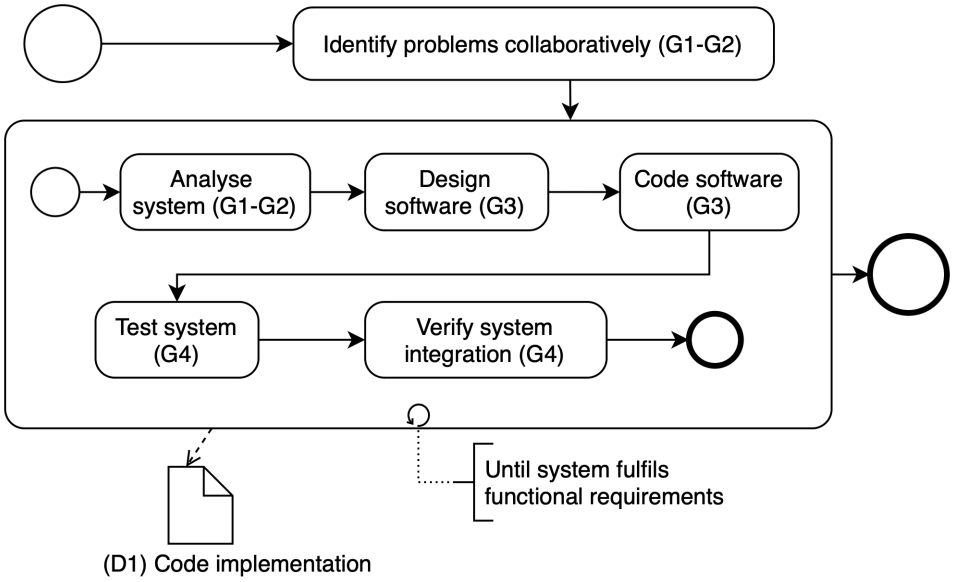


Figure 3.1: BPMN diagram of the System implementation process

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 [45] text editor is development tool of choice in combination with Conquer of Completion (CoC) [46] providing language-aware autocompletion and rust-analyzer [47] access for on-code compiler errors.

- **Libraries and dependencies**: for simpler development and test reproducibility, the environment will be set in a Docker container [48].
- **Code versioning and shared development**: GitHub [49] 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 Spark-based 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 Spark-based access to Apache Hudi compared a the delta-rs [31] 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:

- **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

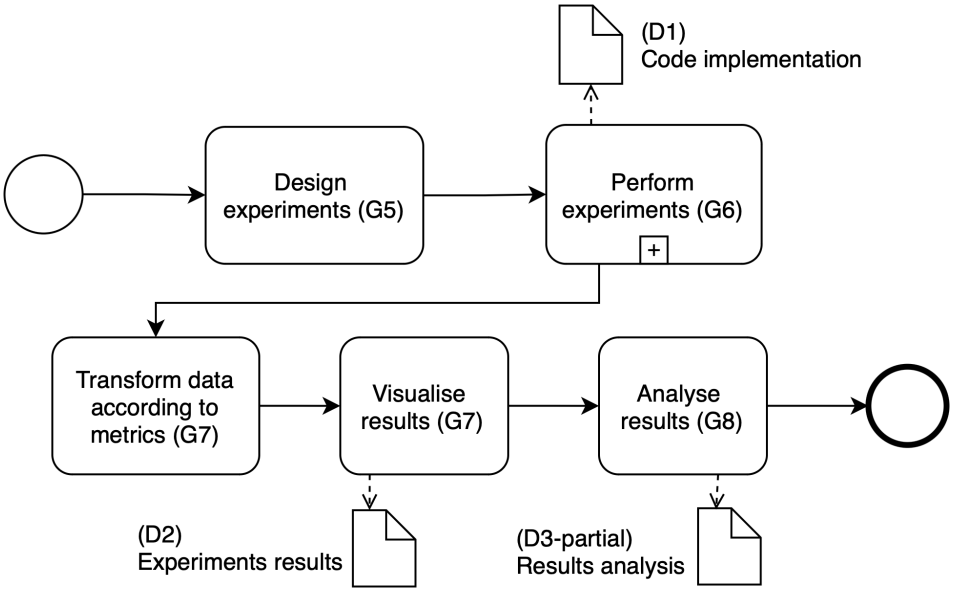


Figure 3.2: BPMN diagram of the System evaluation process

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[50]. 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 [51]. The data coming from this benchmark suite was used as it provides a recognized standard for data storage systems [52], and it has already been used in similar related work [8, 53]. Any part of the data can be generated using the TPC-H data generation tool [54].

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 legacy Spark-based 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 Spark 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.

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.

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 Spark pipeline the contribution of different parts of the process will be

measured: namely the upload time and materialization time, dichotomy explained in Section

Ref here time breakdown expl.

. 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.

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 [31] library. This step outlines the library structure (colloquially/informally defined as "crate") divided into sublibraries (colloquially/informally 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.

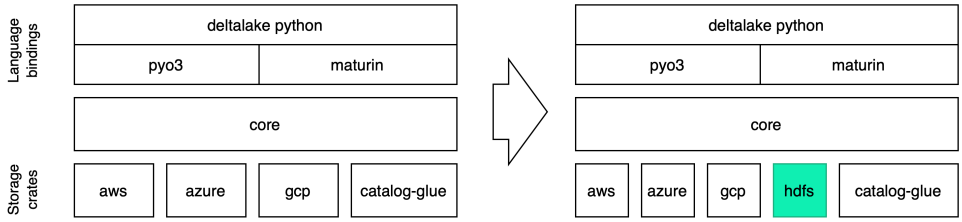


Figure 4.1: delta-rs library (v. 0.15.x) before and after adding HDFS support

The library uses an external interface called object-store defined in the Arrow-rs library [55]. 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 trivia solution was found.

2. The HDFS support had JVM dependencies which weighted a lot on the environment required on delta-rs users (i.e. having Java installed), and this was considered a requirement that would some users from using 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 this past support for HDFS in the delta-rs library, 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 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

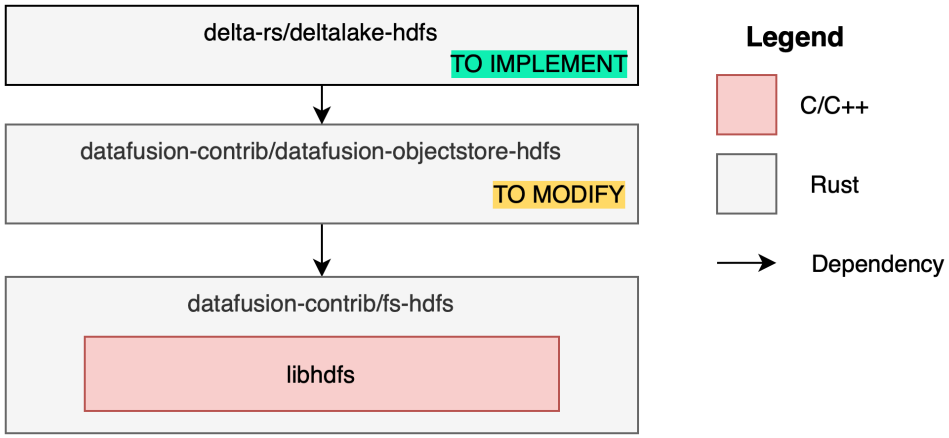


Figure 4.2: Architecture of the first implementation approach

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 [56]. 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.

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

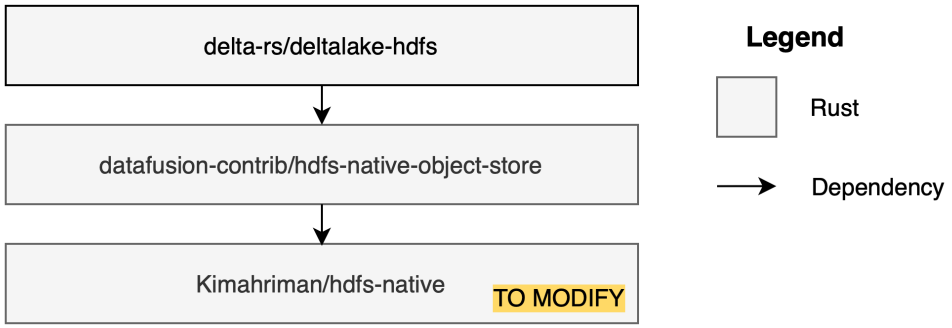


Figure 4.3: Architecture of the final implementation approach

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 [57].

All changes were applied to copies (in Git slang called "forks") 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 [58, 59, 60].

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 [61].

The usage of the delta-rs library is explained in detail in the delta-rs documentation [31], 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 in 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 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 be still be visioned in the Appendix

Add ref to appendix

where for each experiment a histogram and a table containing the 95% confidence interval was 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

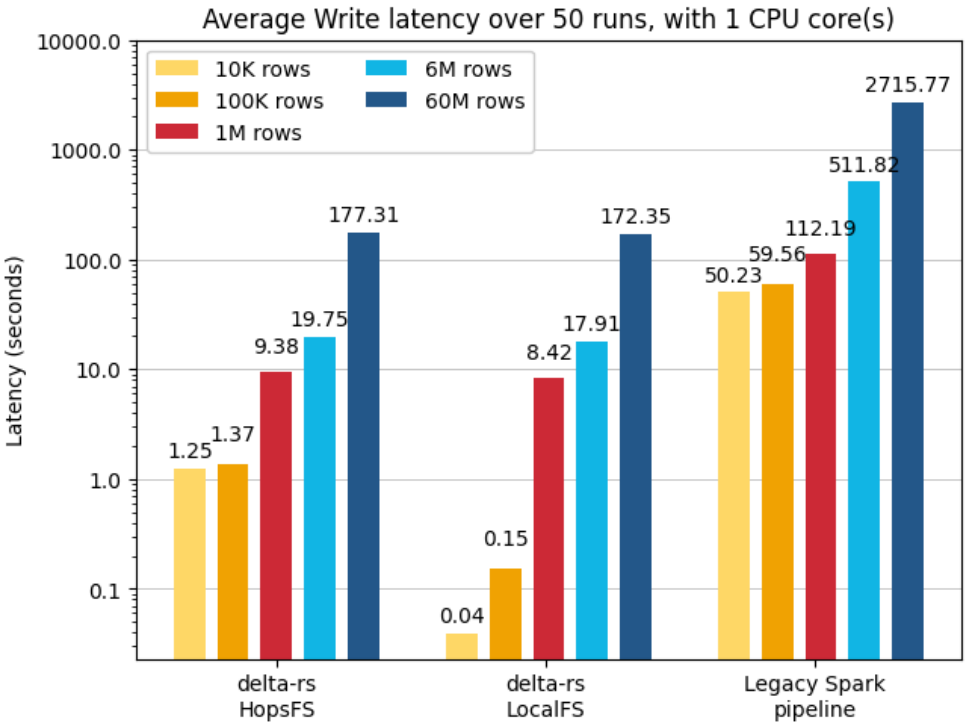
Figure 5.1 and Figure 5.2 show the write latency and throughput respectively of the write operations performed on three different systems defined in Section 3.2.5 when writing the five different tables defined in Section 3.2.4. Both histograms, i.e. Figures 5.1a and 5.2a, report the data from the 1 CPU core experiment while the tables, i.e. Figures 5.1b and 5.2b, 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 on using the delta-rs on HopsFS pipeline in all experiments.

delta-rs on HopsFS vs. Legacy Spark 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 Spark 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 times lower (thr:higher) than the latency measured using the Legacy Spark pipeline.

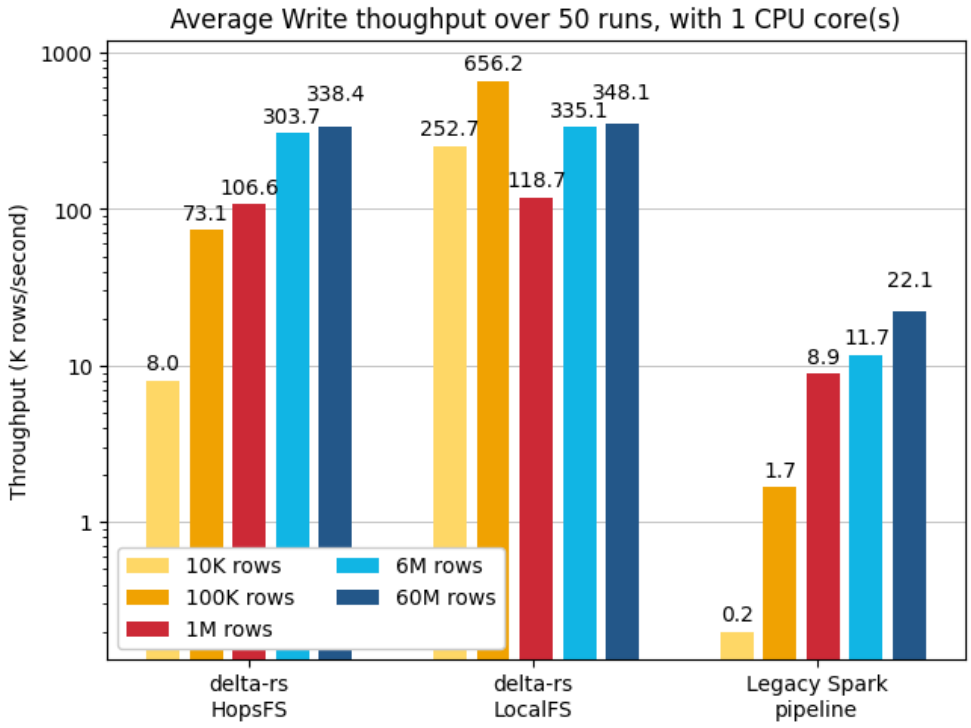


(a) Histogram with write latency experiment results

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	1.250881458553	-0.92	2.75	-9.33
	100K	1.368281275251	4.40	2.34	5.54
	1M	9.381529912744	9.23	10.32	11.52
	6M	19.75469028131	17.54	17.87	20.33
	60M	177.3070700216	24.39	30.01	31.22
delta-rs LocalFS	10K	0.039575374951	-21.88	-15.53	-11.25
	100K	0.152402458205	10.01	13.54	10.45
	1M	8.422528293415	14.69	14.68	14.17
	6M	17.90634508900	14.74	18.71	20.24
	60M	172.3455279090	24.67	29.57	30.38
Legacy Spark	10K	50.22767340044	-0.99	-2.10	-1.99
	100K	59.56187132072	-0.38	0.06	-1.20
	1M	112.1904872782	3.23	3.01	2.50
	6M	511.8169330835	7.51	5.83	7.01
	60M	2715.772857148	13.81	13.61	14.39

(b) Table containing the write latency experiment results compared across multiple CPU configurations

Figure 5.1: Histogram (a) and Table (b) reporting the write latency experiment results also across different CPU configurations



(a) Histogram with write throughput experiment results

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	7.994362640538	-0.91	2.83	-8.53
	100K	73.08438828236	4.60	2.40	5.87
	1M	106.5924224834	10.17	11.51	13.01
	6M	303.7253388718	21.27	21.76	25.52
	60M	338.3959815741	32.26	42.87	45.39
delta-rs LocalFS	10K	252.6823817158	-17.95	-13.44	-10.12
	100K	656.1573952099	11.13	15.66	11.67
	1M	118.7291945082	17.22	17.21	16.51
	6M	335.0767546463	17.29	23.02	25.38
	60M	348.1378410449	32.76	41.99	43.65
Legacy Spark	10K	0.199093434415	-0.98	-2.06	-1.95
	100K	1.678926430325	-0.38	0.06	-1.19
	1M	8.913411682753	3.34	3.10	2.57
	6M	11.72294156790	8.12	6.19	7.54
	60M	22.09315843262	16.02	15.76	16.81

(b) Table containing the write throughput experiment results compared across multiple CPU configurations

Figure 5.2: Histogram (a) and Table (b) reporting the write throughput experiment results also across different CPU configurations

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 Spark 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 on the 2 CPU cores experiments, remaining on similar improvements even with more CPU cores.

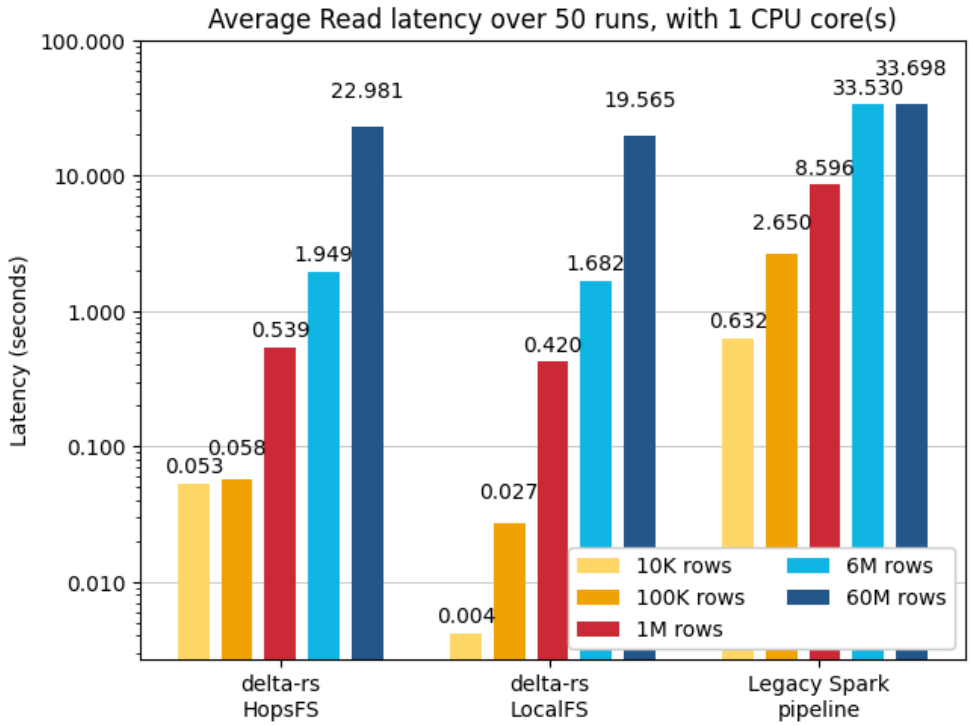
Considering the Legacy Spark pipeline, experiments with more CPU cores did not decrease (thr:increase) the latency (thr:throughput) by more than 7% (thr: 8%) with the exception of the largest table (60M rows). This table benefitted from a latency (thr:throughput) decrease (thr:increase) of around 14%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

5.1.2 Reading Experiments

Figure 5.3 and Figure 5.4 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 5.3a and 5.4a, report the data from the 1 CPU core experiment while the tables, i.e. Figures 5.3b and 5.4b, 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 in the experiment with the smallest table (10K rows). On the other hand, the latency (thr:throughput)

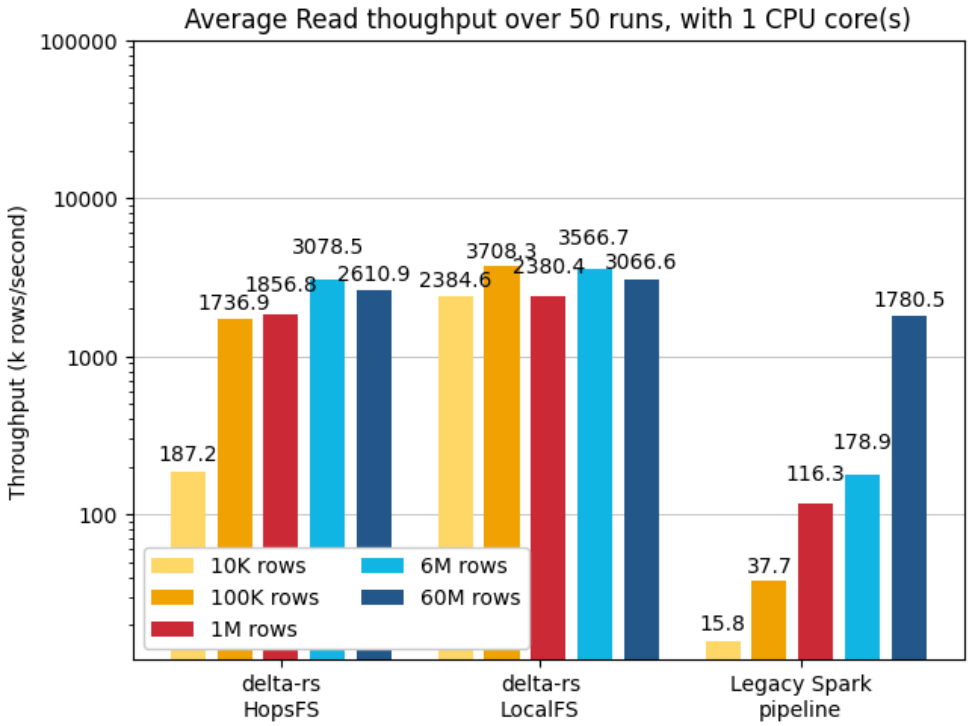


(a) Histogram with read latency experiment results

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	0.053427784961	22.65	18.84	18.95
	100K	0.057573573334	1.15	3.76	5.19
	1M	0.538551778791	56.53	65.00	67.71
	6M	1.948992908833	53.40	72.74	74.48
	60M	22.98065653875	50.34	75.72	87.20
delta-rs LocalFS	10K	0.004193598319	31.48	35.91	29.66
	100K	0.026966840860	51.54	65.76	64.84
	1M	0.420096789910	52.45	78.64	89.75
	6M	1.682239274377	55.56	77.99	89.57
	60M	19.56547400704	51.72	75.41	88.32
Legacy Spark	10K	0.631598152258	1.06	-0.67	0.67
	100K	2.650107346201	-0.50	0.39	-0.46
	1M	8.596367535570	-0.24	-1.81	2.89
	6M	33.52964549822	0.46	0.23	0.30
	60M	33.69772492026	0.16	0.13	1.64

(b) Table containing the read latency experiment results compared across multiple CPU configurations

Figure 5.3: Histogram (a) and Table (b) reporting the read latency experiment results also across different CPU configurations



(a) Histogram with read throughput experiment results

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	187.1685305156	29.28	23.21	23.38
	100K	1736.907998044	1.17	3.90	5.48
	1M	1856.831672237	130.02	185.74	209.69
	6M	3078.512996536	114.57	266.87	291.92
	60M	2610.891464254	101.35	311.83	681.03
delta-rs LocalFS	10K	2384.586991466	45.94	56.04	42.18
	100K	3708.257875522	106.37	192.07	184.38
	1M	2380.403811732	110.28	368.24	875.15
	6M	3566.674545879	125.01	354.40	858.64
	60M	3066.626445053	107.11	306.75	756.07
Legacy Spark	10K	15.83285189203	1.07	-0.67	0.67
	100K	37.73432051472	-0.50	0.39	-0.45
	1M	116.3282044261	-0.24	-1.78	2.98
	6M	178.9461209876	0.46	0.23	0.30
	60M	1780.535633843	0.16	0.13	1.67

(b) Table containing the read throughput experiment results compared across multiple CPU configurations

Figure 5.4: Histogram (a) and Table (b) reporting the read throughput experiment results also across different CPU configurations

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 on using the delta-rs on HopsFS pipeline in all experiments.

delta-rs on HopsFS vs. Legacy Spark 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 Spark 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 a 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.

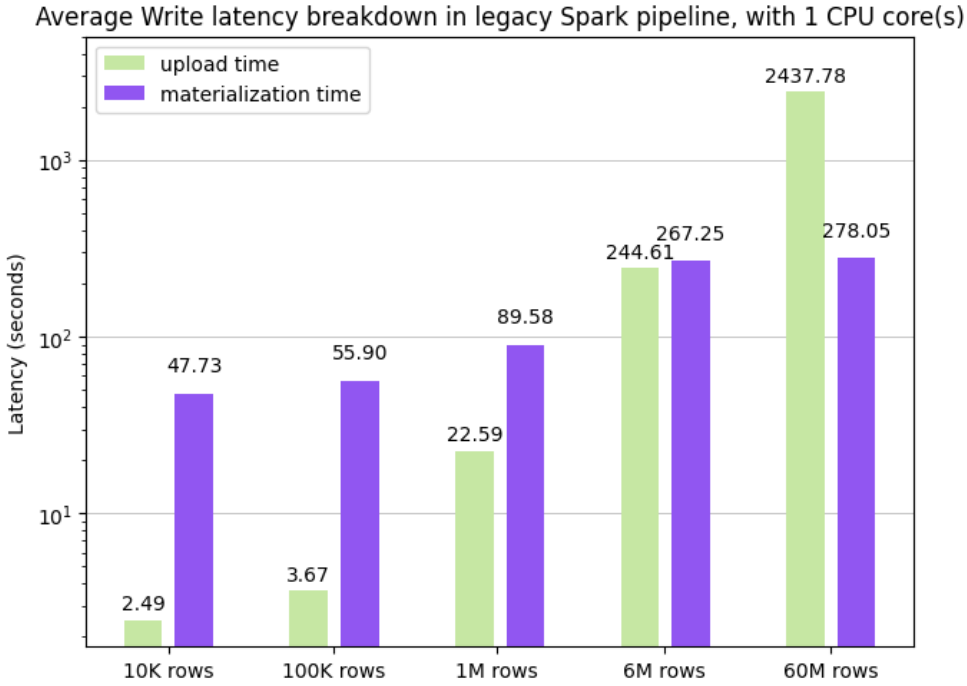
Considering the Legacy Spark pipeline, experiments with more CPU cores did not decrease (thr:increase) the latency (thr:throughput) by more than 2% (thr: 8%). Histograms comparing the three pipelines, look radically different in experiments with more CPU cores, due to how delta-rs scales with the increase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

Add ref. to appendix

.

5.1.3 Legacy Spark pipeline write latency breakdown

Figure 5.5 shows the write latency breakdown of the Legacy Spark pipeline into upload time and materialization time, the different steps of the Legacy Spark pipeline as explained in Section



(a) Histogram displaying the contributions to the write latency of the upload and materialization steps in the legacy Spark pipeline

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	2.48653	47.726	3.99	-1.27	4.10	-2.47	4.22	-2.35
100K	3.66844	55.900	6.37	-0.78	5.55	-0.27	6.25	-1.69
1M	22.5934	89.575	17.52	-0.39	14.89	-0.01	16.51	-1.05
6M	244.612	267.24	15.83	-0.10	13.46	-1.15	15.10	-0.38
60M	2437.78	278.05	15.33	0.39	15.15	0.16	15.94	0.82

(b) Table showing the contributions to the write latency of the upload and materialization steps in the legacy Spark pipeline and it change at different CPU configurations

Figure 5.5: Histogram (a) and Table (b) reporting the contributions to the write latency of the upload and materialization steps in the legacy Spark pipeline and it change at different CPU configurations

Insert ref. to upload materialization expl.

. The breakdown is proposed for all the five different tables defined in Section 3.2.4. Figure 5.5a reports the data from the 1 CPU core experiment while Figure 5.5b 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 materialize contribution to the write latency, making the upload time the 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 an higher number of CPU cores in particular with larger tables (15% latency decrease) 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 and analyses the results presented in Section 5.1. Contrarily to Section 5.1 where results are grouped by experiment type (reading or writing), here each subsection is formed around an observation based on the results that is then motivated and explored, considering all relevant experiments.

5.2.1 delta-rs performance on HopsFS are similar to the library's performance on the LocalFS

The delta-rs on HopsFS pipeline performs similarly (same significant figure) to the delta-rs on LocalFS in all experiments performed on tables from 1M to 60M rows. This suggests that the developed solution takes advantage of the library's full potential. Smaller tables, i.e. 10K and 100K rows, still perform better on the LocalFS (around 10 times better), but this should be caused by limitations given by the different nature of HopsFS, i.e. a distributed system.

5.2.2 delta-rs outperforms the Legacy Spark pipeline in every experiment

The delta-rs on HopsFS pipeline outperforms the Legacy Spark pipeline by a factor of ten or more in write operations, and by a lower factor in read operations. While the trends suggests that with higher table dimensions there might be a cutoff between using the delta-rs pipeline and the legacy Spark pipeline, the implemented system greatly outperforms the legacy Spark system in the designed experiments.

This suggests that in the use case defined in Section 3.2.3, which lead to the experiments design, delta-rs should be adopted as preferred solution over the legacy Spark pipeline.

5.2.3 delta-rs scales well with increased CPU cores

Both delta-rs pipelines scale considerably as the CPU cores increase, by about 30% (latency and throughput) in write operations and linearly for read operations (inverse for latency). On the other hand the legacy Spark pipeline scales poorly, with limited improvements with more CPU cores.

This difference between the two systems is caused by the different architecture of the pipelines. While delta-rs performs operations (read or write) on the local machine, the legacy Spark pipeline uses the local machine only to upload the table in Kafka as messages (this is why the upload time improves with more CPU cores), while it uses other computation present on a Spark cluster to perform the read or write. This means that to improve performances there is the need to give more resources to the Spark cluster, which is an aspect not controlled in these experiments (reproducing a typical scenario in the case of a Hopsworks client's deployment).

5.2.4 The upload time in the legacy Spark pipeline becomes the bottleneck as table size increases

The two steps of the legacy Spark application writing a table show as the materialization time, i.e. when Spark computes, have a great contribution on the overall latency with smaller tables, but do not grow linearly with the table sizes. On the other hand, the upload time scales linearly with the table size from the 1M table, effectively composing 95% of the overall time for the largest table (60M rows). This is caused by how the upload process was implemented in the legacy Spark pipeline, which iterates on each of a tables rows to upload them as Kafka messages.

This highlights as the Legacy Spark pipeline while it might perform better than the implemented system (delta-rs on HopsFS) with larger tables (more than 60 M rows), it is limited in the write operation by the upload operation which scales linearly with a table size. Using a different approach on how data is uploaded in Kafka might improve performances on the legacy Spark pipeline.

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. [Page 1.]

[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] 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 and 3.]

[6] 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 and 4.]

[7] 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.]

[8] 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, and 26.]

[9] 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.]

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

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

[12] 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.]

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

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

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

[16] "Hopsworks - Batch and Real-time ML Platform," https://www.hopsworks.ai/, 2024. [Pages 2 and 5.]

[17] 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.]

[18] "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.]

[19] 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.]

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

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

[22] P. Rajaperumal, "Uber Engineering's Incremental Processing Framework on Hadoop," https://www.uber.com/blog/hoodie/, Mar. 2017. [Page 3.]

[23] 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. [Page 3.]

[24] 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 3.]

[25] D. Borthakur, "The Hadoop Distributed File System: Architecture and Design," 2005. [Pages 3 and 15.]

[26] 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 3.]

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

[28] P. Carbone, A. Katsifodimos, S. Ewen, V. Markl, S. Haridi, and K. Tzoumas, "Apache Flink™: Stream and Batch Processing in a Single Engine." [Page 4.]

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

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

[31] "Delta-io/delta-rs," Delta Lake, May 2024. [Pages 5, 6, 7, 23, 33, and 36.]

[32] 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 and 9.]

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

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

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

[36] 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 8.]

[37] 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 8.]

[38] "| Sustainable Development," https://sdgs.un.org/. [Page 8.]

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

[40] 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.]

[41] "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.]

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

[43] "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.]

[44] M. L. Despa, "Comparative study on software development methodologies." *Database Systems Journal*, vol. 5, no. 3, 2014. [Page 20.]

[45] "Welcome home : Vim online," https://www.vim.org/. [Page 22.]

[46] "Neoclide/coc.nvim," Neoclide, Sep. 2024. [Page 22.]

[47] H. Fann, "Fannheyward/coc-rust-analyzer," Sep. 2024. [Page 22.]

[48] "Docker Build," https://docs.docker.com/build/, 12:14:28 +0200 +0200. [Page 23.]

[49] "GitHub," https://github.com. [Page 23.]

[50] "TPC-H Homepage," https://www.tpc.org/tpch/. [Page 26.]

[51] T. P. P. C. (TPC), "TPC-H_v3.0.1.pdf," 1993/2022. [Page 26.]

[52] 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 26.]

[53] 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 26.]

[54] "TPC Current Specs," https://www.tpc.org/tpc_documents_current_versions/current_specific [Page 26.]

[55] "Arrow-rs/object_store/README.md at master · apache/arrow-rs," https://github.com/apache/arrow-rs/blob/master/object_store/README.md. [Page 34.]

[56] A. Binford, "Kimahriman/hdfs-native," Aug. 2024. [Page 35.]

[57] “Rustls/tokio-rustls: Async TLS for the Tokio runtime,” https://github.com/rustls/tokio-rustls. [Page 36.]

[58] G. Manfredi, “Silemo/hdfs-native,” Aug. 2024. [Page 36.]

[59] ——, “Silemo/hdfs-native-object-store,” Aug. 2024. [Page 36.]

[60] ——, “Silemo/delta-rs,” Sep. 2024. [Page 36.]

[61] “Delta-rs/python at main · Silemo/delta-rs,” https://github.com/Silemo/delta-rs/tree/main/python. [Page 36.]