Language-Agnostic Integrated Queries in a Polyglot Language Runtime System

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera Italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Filippo Schiavio

under the supervision of
Prof. Walter Binder

September 2022

Dissertation Committee

Prof. Shigeru Chiba	The University of Tokyo, Japan
Prof. Matthias Hauswirth	Università della Svizzera italiana, Switzerland
Prof. Hidehiko Masuhara	Tokyo Institute of Technology, Japan
Prof. Cesare Pautasso	Università della Svizzera italiana, Switzerland
Prof. Heiko Schuldt	University of Basel, Switzerland

Dissertation accepted on 14 September 2022

Research Advisor	PhD Program Director
Prof. Walter Binder	**Prof. Stefan Wolf**

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Filippo Schiavio
Lugano, 14 September 2022

Abstract

Language-integrated query (LINQ) frameworks offer a convenient programming abstraction for processing in-memory collections of data, allowing developers to concisely express declarative queries using general-purpose programming languages. Existing LINQ frameworks rely on the type system of statically typed languages such as C♯ or Java to perform query compilation and execution. As a consequence of this design, they do not support dynamically typed languages such as Python, R, or JavaScript. Such languages are however widely used for developing data-processing applications.

In this dissertation, we propose a new approach to query execution based on query interpretation and just-in-time compilation. We introduce DynQ, a novel query engine which bridges the gap between dynamically typed languages and LINQ frameworks by leveraging just-in-time compilation.

From the user prospective, DynQ is a data-processing library which offers SQL and a fluent API as query languages. Internally, DynQ is language-agnostic, since, by leveraging a polyglot language runtime, it brings the LINQ features to multiple languages without requiring one to implement query operators in multiple languages. Moreover, DynQ can execute queries combining data from multiple sources, namely in-memory object collections as well as on-file data and external database systems.

DynQ offers efficient query execution for different kinds of workloads by implementing a hybrid interpreted-compiled execution model. Our approach allows executing queries on small datasets through interpretation, without incurring the overhead of query compilation. On the other hand, DynQ leverages just-in-time compilation to speed up the execution of long-running queries. Moreover, DynQ implements reusable compiled queries, an efficient code cache which allows reusing the same dynamically compiled code for multiple related queries. In this way, DynQ can optimize high-throughput workloads based on a fluent API, i.e., applications which use data-processing libraries mostly for executing many queries on small datasets, such as e.g. in micro-services, as well as applications which make use of data-processing libraries to perform repetitive queries.

Our evaluation of DynQ shows performance comparable with equivalent hand-optimized code, and in line with common data-processing libraries and embedded databases, making DynQ an appealing query engine for standalone analytics applications and for data-intensive server-side workloads. Moreover, thanks to reusable compiled queries, DynQ can also speed up applications that heavily use data-processing libraries on small datasets via a fluent API.

Contents

Chapter 1

Introduction

This chapter first motivates the need for integrating LINQ into dynamically typed languages (Section 1.1). Next, it presents the goals of this dissertation and its main related challenges (Section 1.2). Then, it summarizes the contributions of this dissertation (Section 1.3). Finally, it outlines the structure of the rest of this dissertation (Section 1.4).

1.1 Motivations

In modern data processing, the boundary between *where* data is located and *who* is responsible for processing it has become very blurry. Data lakes [29] and emerging machine-learning frameworks, such as TensorFlow [108], make it very practical for developers to implement complex data-processing applications directly "in the language" (i.e., in Python or JavaScript), rather than resorting to "external" runtime systems such as traditional relational database management systems (RDBMSs). Such an approach is facilitated by the fact that many programming languages are equipped with built-in or third-party libraries for processing in-memory collections (e.g., arrays of objects). Well-known examples of such libraries are the Microsoft LINQ-to-Objects framework [65] (which targets .NET languages, e.g., C♯) and the Java Stream API [79]. Microsoft's implementation of LINQ not only allows developers to query in-memory collections, but it can be extended with data-source providers [64] (e.g., LINQ-to-SQL and LINQ-to-XML) that allow developers to execute federated queries (i.e., queries that process data from multiple sources). Many systems with similar features have been proposed (e.g., Apache Spark SQL [2]).

Despite the many benefits it offers, LINQ support is currently missing in popular dynamically typed languages, i.e., languages for which the type of a variable

is checked at runtime, such as Python or JavaScript. However, such languages are very popular [99], as an example, JavaScript and Node.JS are widely used to implement data-intensive server-side applications [103]. Moreover, a recent survey [9] shows that the most used languages among professional data scientists are Python, SQL, and R. Consequently, we believe that also data scientists would benefit from LINQ frameworks in data-analytics applications.

While bringing LINQ technology to dynamically typed languages is the main motivation for the work presented in this dissertation, the underlying ideas and techniques are not limited to dynamically typed languages but are applicable in the context of statically typed languages, too. Indeed, as we will further discuss in Section 4.4, our system is able to execute queries without resorting to any static type information, i.e., based only on type information available at runtime. However, whenever static type information is available, e.g., the data schema of an R data frame, our system is able to access and leverage such information to further optimize query execution.

Finally, besides the mentioned benefits of bringing LINQ to dynamically typed languages, achieving efficient query processing in more than one language has additional software-engineering advantages for language implementors. Indeed, different languages have distinct API for data processing, often supporting only a subset of what SQL offers, and language runtimes provide ad-hoc implementations of those API. As a consequence, there are often replicated functionalities across those runtimes. In contrast, our approach integrates a single data-processing module with all the supported languages. It supports the full SQL query language and implements dedicated data-processing optimizations, while at the same time it is designed with extensibility in mind, i.e., integrating new data sources is straightforward. Hence, in addition to the benefits that our system offers to end users, it also opens up the opportunity to simplify and optimize the implementation of the data-processing API in different languages.

1.2 Goals and Challenges

The goal of our research is summarized in the following *thesis statement*:

> LINQ can be efficiently integrated into multiple dynamically typed programming languages, requiring no database schema, enabling queries directly on data structures in the host language.

In particular, our goal is to bring the benefits of LINQ systems from statically typed programming languages to dynamically typed ones, enabling queries

on collections of dynamic objects, requiring neither the introduction of a data schema nor a data-ingestion phase, i.e., the process of copying the data into a managed memory space using a physical data representation suitable for being processed by the LINQ engine.

Our first design goal for bringing LINQ into dynamically typed languages is the modularity of the implementation of the query operators in terms of *language independence*. Indeed, a LINQ system for dynamically typed languages can execute queries on collections of objects from multiple programming languages. Similarly, extending such a system by implementing new optimizations and integrating new data sources and query operators should impact only their respective components, i.e., their implementation should be language independent.

Our second goal is *efficient query processing*, in particular, query execution from a LINQ system on a collection of objects from a dynamically typed language should be as efficient as an equivalent hand-optimized application written in the same language. Moreover, query execution should be efficient both for analytical workloads on large datasets (e.g., database workloads), as well as for high-throughput workloads, where many queries are executed on small datasets (e.g., data-intensive server-side applications).

LINQ systems have been studied from a theoretical point of view [38, 15], and several optimization techniques have been proposed [73, 74, 59]. However, the proposed solutions focus on statically typed languages, where type information is known before program execution. The flexibility of dynamically typed languages imposes additional challenges compared with LINQ engines that process data of known types, as the types of the objects in a processed collection (e.g., an array) can be different from each other, so the engine has to take into account, e.g., that a property may be missing in an object or that it may have a different type than a previously processed object.

During the last years, it has been shown by many researches [75, 52, 106, 91] that efficient query execution can be obtained by compiling SQL queries into machine code. However, implementing a language-agnostic LINQ engine based on query compilation imposes additional challenges, since the engine must be able to emit different code depending on the internal implementation of the processed data structure, for example, the engine can emit offset-based machine code when accessing R data-frames, or hash-lookup-based access code when reading data from JavaScript (map-like) objects.

1.3 Contributions

To enable our goal of designing and implementing a LINQ system for multiple dynamically typed programming languages, our research makes the following contributions:

- We describe DynQ[1], a language-agnostic LINQ engine targeting dynamically typed languages. DynQ can execute queries on dynamically typed object collections (e.g., JavaScript or R objects) as well as on file data (e.g., JSON files) and other data sources. DynQ is exposed to users by means of a language-agnostic API, and is capable of executing queries on different object representations. DynQ does not require defining any data schema (neither provided nor inferred) for executing queries on dynamic collections, as it is able to specialize query operators on the data types encountered during query execution. However, at the same time DynQ can further speed up query execution by exploiting a schema when such information is available. Moreover, DynQ can efficiently integrate user-defined functions (UDFs) within query execution.

- We describe DynQ's approach to query compilation, which relies on self-optimizing abstract syntax tree (AST) interpreters and dynamic speculative optimizations. In contrast to many engines based on query compilation, DynQ does not need to generate machine code before executing a query. Query execution in DynQ begins as soon as the AST nodes have been instantiated, since the execution starts by interpreting those nodes. Such an approach to query execution allows DynQ to avoid compiling queries if the underlying dataset is small. On the other hand, DynQ is highly optimized and benefits from just-in-time (JIT) compilation to speed up query execution on long-running queries.

- We address the problem of optimizing high-throughput workloads, where many queries are executed on small batches of datasets. Indeed, using the DynQ execution model based on interpretation and subsequent compilation for large datasets, the application could end up executing most of those small queries through interpretation. To overcome this performance issue, we introduce *reusable compiled queries*, a novel approach to query execution that relies on an efficient and flexible cache of compiled queries. Such an approach allows the query engine to automatically detect the exe-

[1] DynQ is available as open source project at `https://github.com/usi-dag/DynQ-VLDB`.

cution of a query which is similar to a previously executed one and to reuse the previously compiled code to execute the current query.

- We evaluate DynQ on workloads designed for both databases and programming languages. Our evaluation shows that DynQ performance is comparable with hand-optimized implementations of the same queries and outperforms implementations based on built-in or third-party data-processing libraries in most of the workloads. In particular, our evaluation on relational data shows that the flexibility of DynQ in accessing data in different formats does not impair query execution performance with respect to query engines which known the data schema. Moreover, our evaluation on programming-language workloads highlights the ability of DynQ to efficiently process dynamic objects with unknown schema.

1.4 Dissertation Outline

This dissertation is structured as follows.

- Chapter 2 introduces relevant background information related to our approach, i.e., the Microsoft LINQ implementation targeting .NET languages, existing approaches to query execution, and the GraalVM platform with Truffle, its language-implementation framework, on top of which DynQ has been built.

- Chapter 3 discusses the state-of-the art in the context of our work, i.e., query compilation, embedded databases, the support of user-defined functions in database systems, dynamic optimizations in the context of data-processing and the execution of federated queries.

- Chapter 4 presents DynQ, our novel LINQ framework for dynamically typed languages. In particular, it describes the architecture of DynQ, its approach to query compilation, how DynQ allows extending query execution to new data source through DynQ providers, the language-specific type specialization and conversions, and finally how fluent API are implemented in DynQ.

- Chapter 5 introduces *reusable compiled queries*, a technique which improves performance of high-throughput workloads by allowing the reuse of the same compiled code to execute similar queries relying on a cache of compiled queries.

- Chapter 6 evaluates the performance of DynQ on two programming languages, namely R and JavaScript, on workloads designed for databases and programming languages.
- Chapter 7 concludes the dissertation and outlines future research directions inspired by this work.

Chapter 2

Background

This section gives an overview of the .NET implementation of LINQ and discusses its execution model as well as improvements proposed in the research literature. Then, it introduces the GraalVM platform [118] and the Truffle [116] framework that were used for implementing DynQ.

2.1 Language-integrated Queries

LINQ was first introduced in Microsoft .NET 3.5 to extend the C♯ language with an SQL-like *query comprehension* syntax and a set of query operators [8]. The following is an example of a LINQ query:

```
IEnumerable<int> xs = ...;
var evenSquares = from x in xs
                  where x % 2 == 0
                  select x * x;
```

LINQ implements a lazy evaluation strategy by converting query operators to iterators, a so-called *pull-based* model [97], i.e., each operator pulls the next row from its source operator. In the example query, the `where` and `select` clauses in the query comprehension are de-sugared into calls to the methods `Where` and `Select` defined in the `IEnumerable` interface.

Another important feature of LINQ is its extensibility to new data formats. LINQ can execute queries not only on in-memory object collections, but also on any data type that extends the generic types `IEnumerable` or `IQueryable`. This great flexibility is obtained through so-called LINQ providers, i.e., data-source specific implementations of the mentioned generic types. Relevant examples of

LINQ providers are LINQ-to-XML (that queries XML documents) and LINQ-to-SQL, which converts query expressions into SQL queries and sends them to an external DBMS. LINQ providers can be categorized based on where the query computation actually takes place, namely those that execute in the managed runtime of the application process, and those that delegate the query processing to an external system. Examples of the former are LINQ-to-Objects (for in-memory collections) and LINQ-to-XML (that query XML documents), whilst the main example of the latter is LINQ-to-SQL, which converts query expressions into SQL queries and sends them to an external DBMS.

2.2 Query Execution Models

The C♯ implementation of LINQ executes queries by leveraging the pull-based model, which shares many similarities with the Volcano [35] query execution model in use by many popular relational databases, such as PostgreSQL [84]. It has been shown [60] that the main performance drawbacks of this execution model are virtual calls to the interface methods (e.g., `MoveNext()` and `Current()` in C♯, or `hasNext()` and `next()` in Java), which introduce non-negligible overhead, since they are executed for each input row of each operator in the query plan. In the context of relational databases, the most relevant optimizations for removing such overhead are vectorization [10] and data-centric query compilation [75].

Vectorized query execution, similarly to the Volcano model, uses a pull-based approach. However, the query interpretation overhead is mitigated by leveraging a columnar data representation and batched execution, i.e., instead of evaluating a single data item at a time, query operators work on a vector of items which represents multiple input rows. Data-centric query compilation completely removes the interpretation overhead by generating executable code for a given query. Code generation commonly happens at runtime, using schema and type information to generate code that is specialized for the tables used in a query. Data-centric query compilation adopts a so-called *push-based* model, i.e., each operator pushes a row to its destination operators. Both pull-based and push-based models have been studied from the point of view of compilers and program transformations [73, 74, 54]. Interestingly, it has been shown [97] that, by leveraging compiler optimization techniques such as e.g. loop fusion, method inlining, and scalar replacement, neither model clearly outperforms the other.

A well-known disadvantage of query compilation is the overhead introduced by the compilation itself. For statically typed and compiled languages, query

compilation can be invoked at different times, namely at application compilation time (e.g., in SBQL4J [115]) or during its execution (e.g., in Steno [73]). While the former approach has the advantage of hiding the query compilation cost during the compilation of the application, it imposes serious limitations: the queries must be expressed in the application code, meaning that a system that accepts queries as user input cannot leverage such an approach. In the context of dynamically typed languages (such as e.g. JavaScript or Python), this approach cannot be used in general, because the application source code is directly executed by the runtime. On the other hand, runtime query compilation does not suffer from these limitations, but the compilation cost can shadow the benefits obtained by the optimization passes, in particular for short-running queries. Recent research [57] addresses this issue with an adaptive query compilation model. With such a compilation model, the engine first quickly generates an executable representation of the query and executes it in an interpreter. Then, during query execution, the engine performs adaptive decisions whether to compile a query operator based on execution-time estimations. Such approach is inspired by the implementation of JIT compilers in language VMs.

2.3 GraalVM and the Truffle Framework

DynQ is implemented targeting the GraalVM [118] platform, i.e., a polyglot language runtime compatible with the Java Virtual Machine (JVM). GraalVM is capable of executing programs developed in a variety of popular programming languages, such as Java, JavaScript, Ruby, Python, and R. At its core, GraalVM relies on a state-of-the-art dynamic compiler (called Graal [117]), which brings JIT compilation to all GraalVM languages. Language runtimes for GraalVM (including DynQ) are implemented using the Truffle language implementation framework [116]. Unlike other code-generation frameworks for the JVM or the .NET platform, Truffle does not rely on bytecode generation, but rather on the concept of self-optimizing interpreters [116], i.e., language interpreters that use custom API and data structures enabling explicit and direct interaction with the underlying language VM components (including the JIT compiler). The Graal optimizing compiler has special knowledge of such API, and is capable of generating efficient machine code by means of partial evaluation [47].

In addition to JIT compilation, the Truffle framework provides a so-called language interoperability protocol [36], i.e., a runtime mechanism which allows sharing values among the dynamically typed languages implemented with Truffle. Thanks to these interoperability mechanisms, DynQ can effectively inline

machine code used by GraalVM language runtimes into its own query execution code. For example, DynQ can use the very same machine code used by the GraalVM JavaScript VM to read JavaScript heap-allocated objects, thereby enabling efficient access to in-memory data during SQL query execution. This approach to SQL execution allows DynQ to efficiently exploit runtime information, to benefit from optimizations that are normally used in high-performance language VMs, such as e.g. dynamic loop unrolling [21] and polymorphic inline caching [43].

Chapter 3

Related Work

This section discusses the state-of-the-art in the context of this proposal. Section 3.1 discusses query compilation, which is a technique used also in DynQ. Section 3.2 presents an overview of existing embedded databases with bindings for dynamically typed languages and discusses how they differ from LINQ engines, such as DynQ. Section 3.3 discusses different techniques for executing UDFs within a query engine, which is a crucial feature in a LINQ framework. Finally, Section 3.4 discusses related work about dynamic optimizations in the context of data-processing systems.

3.1 Query Compilation

Query compilation in relational databases dates back to System-R [12] and it has been studied in many research work. Recently, query compilation is increasingly gain interest both in the research community [76, 83, 50, 91, 75, 60, 33, 70, 81, 56, 55] and in industrial systems [30, 22, 14, 114]. In the context of stream libraries and fluent API, Steno [73] exploits query compilation in LINQ for the C♯ language. Nagel et al. [74] further improve LINQ query compilation in C♯ by using more efficient join algorithms and by generating native C code which is able to access C♯ collections that reside on the heap of the managed runtime. OptiQL [105] is a stream library for the Scala language which leverages the Delite [104] framework for generating optimized code. Strymonas [54] is a stream library for Java, Scala, and OCaml. Strymonas leverages the LMS [93] framework for ahead-of-time query compilation for Java and Scala, and MetaOCaml [53] for the OCaml language. Those libraries are designed for statically-typed languages and exploit type information for generating specialized programs during the code generation.

Caching the generated machine code of a compiled query for later reuse has been recently proposed in a PostgreSQL query compiler [82]. However, this approach presents two important drawbacks. First, caching is enabled only when developers make use of prepared statement by marking the variables which need to be bind at query execution time, meaning that the developers have to take care of identifying the queries which are suitable for being reused and writing them as prepared statements. Second, bind variables can be only of raw types, i.e., a prepared statement cannot be parametric for a whole expression tree. On the other hand, with reusable compiled queries, DynQ is able to automatically detect the execution of a query similar to a previously executed one and to reuse the previously compiled code to process the current query. Moreover, reusable compiled queries are parametric for whole expressions, thus broadening their applicability in comparison with prepared statements. Permutable compiled queries [71] also addresses the problem of avoiding recompilation. However, such an approach has been designed for integrating adaptivity in compiled queries, and does not allow reusing the previously compiled code for executing subsequent queries, as done in DynQ with reusable compiled queries.

3.2 Embedded Databases

Due to the popularity of dynamically typed programming languages, embedded database systems such as SQLite [80] often provide bindings for some of them. With such an approach, the database query engine is hosted in the application process, removing the inter-process communication overhead imposed by solutions that adopt an external database system [88]. However, developers cannot use embedded databases to query arbitrary data that resides in the process address space (e.g., an array of JavaScript objects or a file loaded by the application). Instead, using embedded databases, it is usually required to create tables with a data schema, and then traverse the object collection and insert relevant data into such tables, a so-called ingestion phase.

Afterburner [28] is an in-memory, embedded relational database implemented in JavaScript. Afterburner leverages optimized JavaScript data structures (i.e., typed arrays) and generates ASM.js [3] code, which is an optimized subset of JavaScript with only primitive types. Although this design offers very fast query processing, it comes with many limitations. First, it cannot execute queries on arbitrary JavaScript objects, i.e., the data needs to be inserted into a database-managed space before query execution, which introduces overhead, increases the memory footprint, and requires users to provide a data schema. Another

drawback of an ingestion phase is that, if the dataset is already stored in a collection (e.g., in an array), copying the data into a managed memory space may significantly increase the memory footprint. Moreover, Afterburner is designed for relational data of few primitive types (i.e., numbers, dates, and strings), with no support for querying arrays and nested data structures. Finally, our goal is supporting data processing in multiple programming languages. However, the approach proposed in Afterburner cannot be easily replicated in other dynamic languages, since most of them do not offer efficient data structures like typed arrays and an efficient subset of the language to operate on primitive datatypes, like ASM.js.

DuckDB [89] is an embedded database with bindings for multiple dynamically typed languages, i.e., Python, R and JavaScript. Differently from most of the other embedded databases, DuckDB is able to execute queries directly on data structures managed by a dynamic language, in particular Python and R data frames. However, both Python and R data frames are implemented with a columnar data structure composed of typed arrays, and they cannot store heterogeneous objects, such as e.g. a JavaScript map. Indeed, DuckDB cannot execute queries on arbitrary objects, in contrast to DynQ. Hence, DuckDB does not need to face the challenge of dealing with unexpected types during query execution.

3.3 Supporting Polyglot UDFs

Efficient execution of UDFs within data-processing systems is an active research area [92], and there have been proposed two types of approaches to address this problem, namely translating UDFs to relational expressions and embedding language runtimes in the data-processing engine.

The first type of approaches, based on translation from UDF to SQL expression [7, 16] has been largely studied in the context of Froid [90], a component in SQL Server which can translate imperative code without loops into SQL queries. Many other approaches [25, 26, 39] improved Froid by allowing supporting more complex imperative code with loop and recursion, converting it into SQL queries with to recursive common table expressions. However, most of the existing approaches for converting imperative code, particularly in the context of dynamically typed languages, are limited to a small subsets of language features [37]. Moreover, often those translations result in complex queries possibly involving recursion, which are known to be difficult to optimize [90].

Other systems compile UDFs into low-level IR such as LLVM [62]. As an example, Tupleware [19, 20, 18] is a state-of-the-art system that translates Python

UDFs to LLVM. However, being based on ahead of time compilation, Tupleware presents many limitations. First, it only supports UDFs for which types can be inferred statically. Moreover, not all Python types are supported, but only numerical types and there must be no polymorphic types as well as NULL values. Finally, there is no exception support for runtime exception. Similar, also Numba [61], a popular Python compiler for scientific computing, supports a limited subset of Python. Tuplex [100, 101] improves Tupleware allowing exception handling. Tuplex introduces a so-called dual mode execution model, which compiles an optimized fast path for the common case (i.e., the expected object shape) which is inferred with sampling, and falls back on CPython otherwise.

The second type of approaches is based on the idea of integrating a whole language runtime to execute UDFs into a data-processing engine, an approach used by existing systems, e.g., Apache Spark [119]. Simpler approaches of these category [67] send each processed tuple to the language runtime, introducing a high overhead. More complex approaches [94, 87] mitigated such overhead by leveraging batched execution, i.e., invoking a UDF once to process multiple rows. However, all this approaches require to execute UDFs written in a certain language on objects of the same language, therefore limiting the reuse of existing libraries as UDFs.

Recently, a query engine which leverages Truffle for optimizing polyglot UDF execution has been proposed in Babelfish [37]. Babelfish and DynQ share some implementation choices, i.e., leveraging Truffle nodes as representation of the operators in a physical query plan. However, Babelfish's goal is efficiently integrating a polyglot UDF execution within a database system with static type information (i.e., the schema). In particular, Babelfish not only requires a schema definition on the input dataset, but also all expressions involved in the query plan need explicit physical type information. On the other hand, our goal is integrating a query engine in LINQ style within dynamically typed languages. Indeed, DynQ deals with query execution on unknown types both on the datasets and the expressions, e.g., executing queries with UDFs on arrays of JavaScript objects. This flexibility is crucial for embedding LINQ-style queries and a fluent API within a dynamically typed language.

3.4 Dynamic Optimizations

Dynamic optimizations are an extensively studied topic in the context of programming languages and just-in-time compilers [21, 43]. In the context of data-processing systems, dynamic optimizations have been studied for adaptive query

processing [5, 42, 63]. However, those techniques are commonly adopted only on systems based on query interpretation, as it is often argued that it is very challenging to achieve on systems based on query compilation, citing [51]:

"Integrating adaptivity in compiled execution is very hard; the idea of adaptive execution works best in systems that interpret a query – in adaptive systems they can change the way they interpret it during runtime. Vectorized execution is interpreted, and thus amenable for adaptivity. The combination of fine-grained profiling and adaptivity allows VectorWise to make various micro-adaptive decisions."

As we will show throughout this dissertation, adaptivity can be integrated in compiled query engines by leveraging dynamic compilation techniques and deoptimizations [44]. Moreover, speculative optimizations for query compilation based on Truffle have been proposed in our VLDB '20 paper [95] in the context of Spark SQL for optimizing query execution on textual data formats. The proposed optimizations target the leaves of a query plan (i.e., table scans with pushed-down projections and predicates). The speculative optimizations discussed in [95] are complementary to the approach used for implementing DynQ, and such optimizations can be integrated into our DynQ providers for textual data sources.

Besides data processing, the Truffle framework has been successfully adopted for optimizing existing libraries. FAD.js [11] is a runtime library for Node.js which optimizes JSON data access by parsing data lazily and incrementally when the data is actually consumed by the application. FAD.js focuses on optimizing data access, whilst DynQ focuses on data processing. Moreover, the approach described in FAD.js is complementary to our approach and can be synergistic with DynQ, i.e., we could integrate FAD.js in the DynQ JSON provider.

3.5 Execution of Federated Queries

As introduced in Section 2.1, the Microsoft implementation of LINQ allows executing queries not only on object collections, but also on any data format for which a so-called LINQ provider is available, i.e., an interface between the data-processing systems and its data source.

In the context of relational DBMS, orchestration of federated queries is a widely studied topic, pioneered by systems like Garlic [48] and TSIMMIS [13] and later extended in other systems such as Apache Calcite [6], FORWARD [32], and Apache Drill [40]. Federated database systems have gained the attention of the research community and have been extended in the context of polystore systems. By following a recently introduced taxonomy [107], query processing

across multiple data sources can be grouped in the following categories: Federated databases (e.g., Multibase [98]) which are composed of multiple homogeneous data sources offering a single query language. Polyglot systems (e.g., Apache Spark SQL [2]) which are again composed of multiple homogeneous data sources but they support multiple query languages. Multistore systems (e.g., FORWARD [32]) which are composed of multiple heterogeneous data sources and offer a single query language. Polystore systems (e.g., BigDAWG [24, 34], CloudMdsQL [58], Polypheny-DB [113]) which are composed of multiple heterogeneous data sources and offer multiple query languages.

According to the mentioned categories, DynQ belongs to the category of multistore systems, since it allows executing queries on heterogenous data sources and it offers SQL as query language. Internally, DynQ leverages Apache Calcite [6] to orchestrate query execution among different engines.

Chapter 4

DynQ - a Dynamic LINQ Engine

This chapter presents DynQ, our novel LINQ engine which bridges the gap between LINQ frameworks and dynamically typed languages. First, we give a detailed description of DynQ's internals, presenting first the general design of DynQ (Section 4.1), then it's approach to dynamic query compilation (Section 4.2), and its built-in support for third-party data providers (Section 4.3). We also explain how DynQ's architecture facilitates the development of language-specific optimizations (Section 4.4). Finally, we describe the implementation of the fluent API in DynQ (Section 4.5). Unless otherwise specified, the achievements presented in this chapter were published [96] at VLDB '21.

In designing DynQ we focused on the following goals:

- *Language-independence and modularity:* DynQ should be able to execute queries on collection of objects from any language supported by GraalVM. Moreover, integrating new data sources and query operators in DynQ should impact only their respective components, i.e., their implementation should be language-independent.

- *High performance:* Query execution with DynQ from a host language should be as efficient as a hand-optimized application written in the same language.

For creating a language-agnostic query engine, one could follow a canonical compiler design approach, by first implementing a common front-end for the query language (e.g., SQL) with a parser and a common optimizer (e.g., a query planner), and then implementing language-specific back-ends that compile the query plan into source code for any target language. Using this approach, a query engine could achieve query execution performance in line with an equivalent

hand-optimized application written in the target language. However, such an approach would not meet language-independence requirement listed above, since extending the engine with new query operators or data sources would require extending all language-specific back-ends, i.e., the query operator implementations are not language-independent. In the following subsections we describe how we designed DynQ to meet both the mentioned requirements.

4.1 DynQ Architecture

At its core, DynQ is a dynamic query engine for GraalVM that exploits advanced dynamic compilation techniques to optimize query execution. DynQ is exposed to users by means of a language-agnostic API, and is capable of executing queries on any object representation supported by GraalVM languages.

Unlike the popular LINQ implementation for the .NET platform, DynQ does not extend its supported programming languages with a query-comprehension syntax, but rather relies on SQL queries expressed as plain strings. The LINQ query-comprehension syntax allows query validation at program compilation time. However, as already discussed in the literature [46], in a dynamically-typed language, where syntactic validation and type checking take place at runtime, lacking this form of compile-time validation is not an issue. Query-comprehension could be integrated in a dynamically typed language without any form of compile-time validation. However, since one of the main goals of DynQ is language independence, extending the syntax of multiple languages would not be a practical approach. For this reason, and for the sake of a simpler implementation, we opted for expressing SQL queries as strings. Besides SQL, DynQ allows expressing queries using a so-called fluent API, as described in Section 4.5. Since DynQ is currently a prototype, its fluent API do not cover all the operations implemented in .NET LINQ. However, besides offering comprehensions as query language and static type checking, DynQ approach to query execution is compatible with all the features offered by the LINQ implementation for the .NET platform. Implementing the missing operations would require only engineering effort.

Two important differences between DynQ and existing LINQ systems are its dynamic type system and the tight integration with the underlying JIT compiler. The flexibility of dynamic languages imposes additional challenges compared with processing data with a known type, as the engine has to take into account that a value may be missing in an object and that runtime types of the objects in a single collection (e.g., an array) can be different from each other. JIT compilation is crucial in this context, as it allows DynQ to generate machine code that is

specialized for the data types observed at runtime. For example, DynQ can emit offset-based machine code when accessing R data frames, or hash-lookup-based access code when reading data from JavaScript (map-like) dynamic objects.

Close interaction with the platform's JIT compiler is a peculiar feature of DynQ and a key architectural difference w.r.t. other popular language-integrated approaches. Existing systems (e.g., .NET LINQ or Java 8 Streams) do not interact with the underlying language runtime; in these systems, queries are compiled to an intermediate representation (e.g., .NET CLR or Java bytecode) like any other language construct (e.g., Java 8 Streams are converted to plain Java bytecode with virtual method calls and loops). Query compilation to such intermediate representations happens statically, before program execution. At runtime, the language VM might (or might not) generate machine code for a specific query. However, the lack of domain knowledge of the underlying JIT compiler could limit the class and scope of optimizations that the language VM can perform. For example, a language VM might (or might not) decide to inline certain methods into hot method bodies depending on runtime heuristics that have nothing to do with the structure of the actual query being executed.

DynQ, on the contrary, takes a radically different approach as it explicitly *interacts* with the underlying VM's JIT compiler to drive query compilation. In this way, DynQ can effectively propagate its runtime knowledge of any given query to machine code generation, resulting in high performance. As an example, DynQ can effectively *force* the inlining of the predicates of a given query expression into table-scan operators, ensuring efficient data access. Moreover, the tight integration with the language VM's JIT compiler unlocks a class of optimization that are not achievable with existing LINQ-like systems, namely, *dynamic* speculative optimizations: not only can DynQ apply an optimization (e.g., inlining) when it sees potential performance gains, but it can also *de-optimize* the generated machine code when certain runtime assumptions get invalidated, giving the query execution engine the chance to re-profile the code that is being executed, possibly leading to the generation of new machine code that now takes different runtime assumptions into account.

Thanks to its design, DynQ is able to outperform hand-optimized implementations of queries written in dynamic languages. Internally, the type system of DynQ's query engine handles two main types, namely primitive types and structured types. Primitive types include all Java primitive types as well as String and Date. Structured types include arrays and nested data structures, i.e., objects with properties of any of the mentioned types; multiple nesting levels are supported as well. As expressions, DynQ supports logical and arithmetic operators, the SQL `LIKE` function on strings, and the `EXTRACT` function on dates. Moreover,

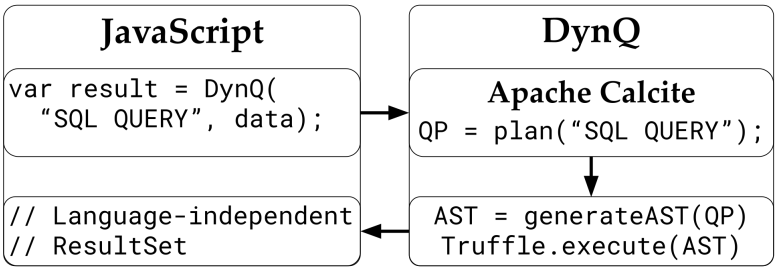

Figure 4.1. High-level query life cycle in DynQ.

DynQ seamlessly supports user-defined functions (UDFs), as it can directly inline code from any GraalVM language into its SQL execution code. In this way, UDFs from any of the GraalVM languages can be called during SQL evaluation with minimal runtime overhead. In particular, it is not required that a UDF is written in the same language as the application that is using DynQ, e.g., it is possible to use DynQ from JavaScript, executing a query with a UDF written in R.

As GraalVM is compatible with Java, DynQ can leverage existing Java-based components to perform SQL query parsing and initial query planning. To this end, our implementation leverages the state-of-the-art SQL query parser and planner Apache Calcite [6]. While using Calcite as an SQL front-end has the notable advantage that DynQ's implementation can focus on runtime query optimization *after* query planning, DynQ's design is not bound to Calcite's API, and other SQL parsers and planners could be used as well.

A high-level overview of the life-cycle of a query executed with DynQ from a dynamic language (JavaScript in the example) is depicted in Figure 4.1. As the figure shows, as soon as a developer has defined a dataset in the form of an object collection (e.g., an array) it is possible to execute an SQL query on the in-memory data. DynQ is invoked from the host dynamic language, passing (as parameters) a string representation of the query and a reference to the input data. DynQ leverages Calcite for parsing and validating the SQL query; if successful, the validated query is converted into an optimized query plan. Then, DynQ traverses the query plan, generating an equivalent executable representation (i.e., Truffle nodes [116]), which is our form of a physical plan.

By generating Truffle nodes, the query preparation phase of DynQ is very efficient, as Truffle nodes are ready to be executed through interpretation. Query execution thus begins by interpreting the Truffle nodes generated by DynQ. As soon as the DynQ runtime detects that the AST (or parts of it) are frequently executed, it delegates the JIT compilation to GraalVM. Dynamic compilation is triggered by DynQ, which also takes possible runtime de-optimizations and re-

compilations into account. Finally, the result of query execution, i.e., a language-independent data structure accessible by any GraalVM language, is returned to the application.

Source-code Metrics Source-code metrics of DynQ are presented in Table 4.1. The parser module contains classes that use Calcite for SQL parsing and query planning, as well as the conversion from a Calcite plan to DynQ Truffle nodes. The expression and query-operator modules contain DynQ Truffle nodes. The polyglot-API module contains classes that implement the API accessible from all languages supported by GraalVM, as well as the implementation of our polyglot result-set.

Table 4.1. DynQ source-code metrics; number of classes and lines of code (LOC) for each module.

	# Classes	# LOC
Parser	37	3700
Expressions	123	5300
Query Operators	116	5800
Polyglot API	43	2500
Total	319	17300

4.2 Query Compilation in DynQ

Query compilation in DynQ uses a push-based approach and takes place by visiting the query plan generated by Calcite and converting it into Truffle nodes. The push-based query execution approach used by DynQ is inspired by the model introduced in LB2 [106]. In this model, each operator produces a result row that is consumed by an executable callback function. Rather than relying on statically generated callback functions, however, DynQ propagates result rows to Truffle nodes. In this way, those nodes can specialize themselves on the actual data types observed at runtime. Internally, DynQ relies on two classes of Truffle nodes, namely (1) Expressions and (2) Query-operators.

Expression nodes represent the supported SQL expressions and UDF functions introduced in Section 4.1. Since DynQ is designed to be a schema-less query engine, each expression node used in a query can have initial unknown input (and output) type. During query execution, Truffle nodes rewrite themselves to

specialized versions capable of handling the actual types observed during query execution. This specialization mechanism is natively supported by the Truffle framework, and allows DynQ to handle type polymorphism in a way analogous to language runtimes, resorting to runtime optimization techniques such as polymorphic inline caches [43]. In this way, an expression can be specialized during query execution to handle multiple data types. Note that, typed expression may be used in DynQ as well, in case the types information of the underlying data source is available, as further discussed in Section 4.4.

Query-operator nodes are responsible for executing SQL operators, eventually producing a concrete result value. DynQ relies on two categories of query-operator nodes, namely *consumer* nodes and *executable* nodes. The corresponding Java interfaces `ExpressionNode`, `ConsumerNode`, and `ExecutableNode` are shown in Figure 4.2.

Intuitively, each query operator (excluding table scans) has its own consumer node, whilst only table-scan and join operators implement an executable node. The main executable node of a query, i.e., the one containing the root operator, takes care of producing the result set for that query.

DynQ generates a query's root executable node by visiting the plan generated by Calcite. In particular, DynQ generates a consumer node C for the currently visited operator O. If O is not a join (i.e., it has only one child), C will consume the rows produced by the child of O. If O is a table scan, DynQ generates an executable node which iterates over a data structure (which acts as a table), invoking the generated chain of consumer nodes for each row.

The implementation of the consumer nodes generated by visiting a join operator depends on the join type. DynQ supports nested-loop joins and hash-joins (possibly with non-equi conditions). In case of nested-loop joins, DynQ creates a left consumer which inserts all rows into a list L, and a right consumer that finds matching pairs of rows by iterating over the elements in L for each row. In case of hash-joins, the left consumer inserts the rows in a hash-map, which is used by the right consumer to find matching pairs.

When the query root operator is not a materializer (e.g., for queries composed of projections and predicates), DynQ adds a custom consumer which fills a list of rows, since DynQ always outputs an array data structure. On the other hand, when the root operator is a sort or an aggregation, DynQ returns the sorted (or aggregated) data, which is already a list of rows.

Note that stateful operators do not need any specific executable node, since they are implemented using the methods defined in the interface `ConsumerNode`, i.e., `consume(row)` and `getResult()`. As an example, if a query has a group-by operator (which is not the root operator in the query plan), its implementation of

```
interface ExecutableNode {
 Object execute();
}

interface ConsumerNode {
 void consume(Object row) throws EndOfExecution;
 Object getResult();
}

interface ExpressionNode {
 Object execute(Object row);
}
```

Figure 4.2. Main interfaces in DynQ.

`consume(row)` updates the internal state (a hash-map) and the implementation of `getResult()`, which is invoked by its source operator once all input tuples have been consumed, sends all tuples from the aggregated hash-map to its destination (a `ConsumerNode`), calling the `consume(row)` method for each aggregated row, and finally returns the value obtained by calling the `getResult()` method on its destination consumer.

Since push engines do not allow terminating the source iteration, i.e., an operator cannot control when data should not be produced anymore by its source operator, DynQ implements early exits for the *limit* query operator by throwing a special `EndOfExecution` exception, which is part of the signature of the method `ConsumerNode.consume(row)`.

Error handling in DynQ. During data-processing on data which is not directly managed by query engine like in a RDBMS, errors due to malformed rows need to be handled by the engine. As an example, Apache Spark [2] allows users to choose among three different strategies in case of malformed rows, i.e., discarding all the malformed rows (permissive mode), throwing an exception at the first malformed row found during data processing (failfast mode), and collecting the malformed rows into the result set (later dumped to a file) for further analysis or data fixing (collect bad records mode).

The latter approach has been extended in the literature in the context of Tuplex [100, 101], a processing model which offers advanced error handling strategies. In particular, in Tuplex, malformed rows are collected during data

processing and stored in the so-called exception set, i.e., a field of the final result set. Such an exception set maps exception types to the list of tuples that generated such exceptions. Besides collecting errors, Tuplex allows developers to define *resolve functions*, i.e., exception handlers which define alternative strategies to clean malformed rows on-the-fly instead of collecting them, e.g., replacing a missing value with a default one. Moreover, Tuplex allows partially replaying pipelines with a so-called incremental exception resolution. In this way, once errors have been collected into the exception set, the pipeline can be modified and re-executed only on the exception set.

Currently, DynQ's approach to error handling is rather simple and works as the permissive mode of Spark, i.e., it discards all the malformed rows during query execution. However, implementing different error-handling strategies like those proposed in Tuplex would only require engineering efforts, since DynQ's approach to query execution is compatible with these approaches to error handling. Moreover, in contrast to Tuplex, where exceptional cases are always executed as slow paths, i.e., relying on the CPython interpreter, DynQ can specialize on exceptional cases if they are common in a certain dataset, leading to possibly more efficient error handling.

Query compilation example. Consider the DynQ query targeting a JavaScript array of objects shown in Figure 4.3. The query execution plan for the example query is composed of a table-scan operator, a predicate operator, and an aggregation operator that counts the number of rows that satisfy the predicate. The AST of Truffle nodes generated by DynQ for the example query is depicted in Figure 4.4. A simplified implementation of the nodes that compose the query is depicted in Figure 4.5 (table scan), Figure 4.6 (predicate), Figure 4.7 (count), Figure 4.8 (less than), Figure 4.9 (read member). As shown in Figure 4.8, the `LessThanNode` node leverages Truffle specializations for implementing the less-than operation. The `LessThanNode` implementation shown in Figure 4.8 presents only the specializations for Int and Date types, because those types are the ones used in the example. The actual implementation contains specializations for all types supported in DynQ as well as their possible combinations (e.g., Int/Double and Double/Int). In particular, DynQ specializations with mixed types respect the implicit type conversion (i.e., type cast) semantics that in a relation database is commonly integrated in the query planner. However, since in general at query planning time types information are not available in DynQ, instead of inserting implicit cast operations in the query plan, the detection of such casts must take place during data processing.

```
var data = [{x: 1, y: 2},
            {x: 2, y: 1},
            {x: Date('2000-01-01'), y: Date('2000-01-02')}];

DynQ.registerTable(data, 'T');
var Q = 'SELECT COUNT(*) FROM T WHERE x < y';
var result = DynQ.execute(Q);
```

Figure 4.3. Example of a DynQ query on a JavaScript array.

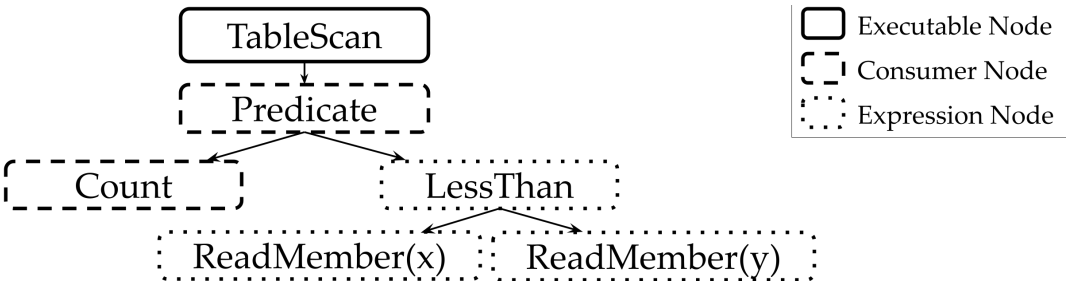

Figure 4.4. AST generated by DynQ for the query in Figure 4.3.

Consider the method `execute(Object)` defined in the class `LessThanNode`. This method first executes the left and right children expression nodes (i.e., property reads in the example query). Then, the method call to `executeSpecialized` (internally) performs a type check for the two arguments (i.e., `fst` and `snd`). If both values have type `int`, the specialization `execute(int, int)` is executed; if they are both dates, the method `execute(LocalDate, LocalDate)` is executed; otherwise, the current tuple is generates an error which will be handled as discussed above.

Consider again the AST generated by DynQ for the example query depicted in Figure 4.4. If the query is executed on an R data frame, DynQ would generate *the same tree*, but the `TableScan` executable node and the `ReadMember` expression nodes would specialize in different ways, depending on the runtime types. The flexible design of DynQ allows reusing the very same query-operator nodes for executing queries on different data structures, like a JavaScript array of objects or an R data frame. Thanks to this design, we achieve all the three goals listed in the beginning of this section. In particular, the extensibility and modularity of our design allow adding new data sources (e.g., a data structure in a dynamic language or an external source like a JSON file) by integrating only the expres-

```
class TableScanNode implements ExecutableNode {
    ConsumerNode consumer;
    PolyglotArray input;

    public Object execute() {
        try {
            for(int i=0; i<input.numElements; i++) {
                Object row = jsArrayElement(input, i);
                consumer.consume(row);
            }
        } catch (EndOfExecution e) {}
        return consumer.getResult();
    }
}
```

Figure 4.5. Simplified DynQ implementation for the table-scan node in the example query in Figure 4.3.

sion nodes which take care of accessing data from such a data source, without requiring any modification to the query-operator nodes.

Dynamic machine-code generation. By implementing DynQ on top of Truffle, DynQ has fine-grained control over Graal, the GraalVM's JIT compiler. Dynamic compilation is triggered based on the runtime profiling information collected during query execution, and the Graal JIT compiler applies (to DynQ queries) all optimizations that are commonly used in dynamic language runtimes. Examples of optimizations applied by Graal include aggressive inlining, loop unrolling, and partial escape analysis. JIT compilation is performed by GraalVM using a configurable number of parallel compiler threads. This leads to short compilation times, as we will further discuss in Section 6.2.

In contrast to many engines based on query compilation, DynQ does not need to generate machine code before executing a query. Query execution in DynQ begins as soon as the Truffle nodes have been instantiated. First, the execution starts by *interpreting* those nodes; during this phase the runtime collects type information for the nodes that leverage Truffle specializations (e.g., `LessThanNode` in the previous example). Then, once the runtime detects that some nodes are frequently executed (e.g., the main loop in `TableScanNode`), it initiates machine-code generation. Once the runtime has collected type information for those rows

```
class PredicateNode implements ConsumerNode {
    ConsumerNode consumer;
    Expression predicate;

    public void consume(Object row) {
        if(predicate.execute(row)) {
            consumer.consume(row);
        }
    }
    public Object getResult() {
        return consumer.getResult();
    }
}
```

Figure 4.6. Simplified DynQ implementation for the predicate node in the example query in Figure 4.3.

```
class CountNode implements ConsumerNode {
    long result = 0;

    public void consume(Object row) {
        result++;
    }
    public Object getResult() {
        return result;
    }
}
```

Figure 4.7. Simplified DynQ implementation for the count node in the example query in Figure 4.3.

```
class LessThanNode implements ExpressionNode {
    ExpressionNode left, right;

    public boolean execute(Object row) {
        Object fst = left.execute(row);
        Object snd = right.execute(row);
        return executeSpecialized(fst, snd);
    }
    @Specialization
    boolean execute(int left, int right) {
        return left < right;
    }
    @Specialization
    boolean execute(LocalDate l, LocalDate r) {
        return l.isBefore(r);
    }
}
```

Figure 4.8. Simplified DynQ implementation for the less-than node in the example query in Figure 4.3.

```
class ReadMember implements ExpressionNode {
    String name;

    public Object execute(Object row) {
        return readJsMember(name, row);
    }
}
```

Figure 4.9. Simplified DynQ implementation for the read-member node in the example query in Figure 4.3.

which have been executed in the interpreter, it speculatively generates machine code assuming that the subsequent rows will have the same types. If such speculative assumptions get invalidated (e.g., because a subsequent row has an unexpected type), the compiled code gets invalidated and the execution falls back to interpreted mode. Then, the runtime can update the collected type information and later re-compile the nodes to machine code accordingly. It is important to note that, even if triggering recompilation has a cost, specializations stabilize quickly [117], typically incurring only minor overhead. Besides type specialization, such a dynamic approach to query compilation is crucial for obtaining low latency in query executions, as we will show in Section 6.2.

By leveraging a state-of-the-art dynamic compiler like Graal, DynQ can selectively compile single components of the query's physical plan. In particular, each table-scan executable node can be selectively compiled to self-contained machine code. Thanks to this approach, a query does not need to be fully compiled to machine code to achieve high performance, since e.g. executing a join operator could lead to the evaluation of one child node in the interpreter (if it has few elements) and another child in compiled machine code.

Figure 4.10 shows the pseudo-code equivalent to the machine code generated by DynQ for the example query of Figure 4.3, once both types in the example are encountered (i.e., both `x` and `y` properties have either type Int or Date). As the figure shows, all the calls to the interface methods are aggressively inlined by the compiler. The operations listed at the beginning of the `while` loop that interact with the host dynamic language (i.e., reading the current array element and its properties `x` and `y`) are inlined by the compiler as well. Moreover, the predicate node is compiled into two `if` statements that check whether in the current row the fields `x` and `y` have one of the expected types. If this is not the case, in general, the generated code would be invalidated as described above, whilst in this specific example, since there is no other specialization in the less-than node, the current row is discarded and the generated code does not need to be invalidated.

4.3 DynQ Providers

As introduced in Section 2.1, LINQ queries are not limited to object collections, instead they can be executed on any data format for which a so-called LINQ provider (i.e., a data-source specific implementation of the LINQ interfaces enumerable and queryable) is available. Such flexibility is an appealing feature for developers, since it allows executing federated queries within the same program-

```
executeMethodAfterJITCompilation() {
    result = 0;
    for(int i = 0; i < input.numElements, i++) {
        row = // read i-th array element
        fst = // read property "x" of row
        snd = // read property "y" of row
        // Type checking for predicate
        if(/* fst and snd are integers */) {
            if(fst < snd) {
                result++;
            }
        }
        else if(/* fst and snd are dates */){
            if(fst.isBefore(snd)) {
                result++;
            }
        }
    }
    return result;
}
```

Figure 4.10. Pseudo-code equivalent to the machine code generated by DynQ, executing the example query in Figure 4.3.

ming model, leaving the complexity of orchestrating different data sources to the system.

In DynQ, we leverage Apache Calcite to federates multiple storage and processing backends. Indeed, besides the query parser and planner, Apache Calcite has another appealing feature for DynQ, namely its flexibility in integrating new data sources by defining specific adapters. A Calcite adapter takes care of representing a data source as tables within a schema, i.e., a representation that can be processed by the query planner. Similarly to LINQ providers, from a query execution point of view, a Calcite adapter takes care of converting the data from a specific source to a Calcite enumerable that can be integrated into the query engine, allowing the execution of federated queries. In this way, similarly to multistore databases, we use a single query language (i.e., SQL) to execute queries that combine data from multiple sources. In implementing DynQ we leveraged Calcite adapters for integrating Java Database Connectivity (JDBC) drivers. As the Calcite ecosystem offers many adapters for other data sources (e.g., Elasticsearch [23], MongoDB [72], Apache Kafka [109]), all these adapters can be integrated in DynQ with minimal effort.

Besides leveraging existing Calcite adapters, DynQ-specific provider can be implemented as well. Implementing a DynQ provider requires implementing a specific table-scan operator, which takes care of iterating over the rows in the input data source, and a data-accessor operator, which takes care of accessing the fields of each row. As an example of DynQ provider, we implemented a JSON data-source provider using Jackson [31], an efficient JSON parser for Java, for accessing fields in JSON objects. This approach can be further extended with more complex parsers that integrate predicate execution during data-scan operations, which is an approach already explored our previous work [95].

As an example of federated query in DynQ, consider a scenario where a developer needs to analyze a web-server log file in JSON format, counting the number of accesses for each user who registered to the website after a specific date, with user registration data however stored in a database. Figure 4.11 shows how such a log analysis can be executed with DynQ. As the figure shows, developers do not have to deal with opening/closing any file or database connection; they only need to provide a file name and configurations for accessing the database (e.g., the URL, credentials, and database name) to DynQ, which takes care of everything else.

The Calcite query planner detects the operators that can be pushed to external data sources. When executing the example query, DynQ sends (to the database) the SQL query with the predicate on the date field and retrieves only user names of the rows matching the predicate. Hence, the operation can be executed more

efficiently (i.e., exploiting database optimizations) and communication overhead is reduced. Currently, DynQ can push down predicates to data sources if they do not involve UDFs. In this case, the predicate operator will be kept in a part of the query plan processed by DynQ. However, the problem of translating UDFs into SQL expressions has already been addressed, e.g., in BOLDR [7], an approach that can be synergistic with our work.

```
var path = 'file://.../log.json';
DynQ.registerJSON('logs', path);
var config = // DB url, credentials, ...
DynQ.registerJDBC('users', config);

var result = DynQ.execute(`
    SELECT users.name, COUNT(*) as count
    FROM users, logs
    WHERE logs.user.id = users.id
    AND users.registration_date > DATE ...
    GROUP BY users.name`);
```

Figure 4.11. Federated query with DynQ.

4.4 Language-specific Expressions

Although GraalVM allows efficient interactions among different languages [36], it may introduce overhead related to data conversion operations. As an example, dates are represented as `LocalDate` instances once shared among different languages, but the internal representation in a specific language may be different, e.g., in JavaScript dates are represented as long values, as the number of milliseconds from the epoch day January 1, 1970-01-01, UTC [27]. As an example, consider the following simple query:

```
SELECT COUNT(*) FROM T WHERE X < DATE '2000-01-01'
```

Suppose DynQ executes such a query (without language-specific expressions) on JavaScript objects, in a first step (before query execution) it would create a `LocalDate` instance for the constant date (2000-01-01), then during predicate evaluation, for each row:

- It would check that the current row contains the field `X` and that it is actually a date instance (this step cannot be avoided in the context of dynamic languages).
- It would convert the JavaScript date into a `LocalDate` instance.
- Finally, it would compare the converted `LocalDate` instance with the constant one (2000-01-01).

On the other hand, evaluating the predicate in JavaScript would require only the first step above (i.e., checking that the field exists and has type date), if so the date comparison is executed using the JavaScript internal representation of dates, that is, a single comparison of two primitive longs, which is of course much more efficient than the steps above.

The reason for those data conversions is that different languages may internally represent the same data type differently, but exposing those types to other languages requires a common representation. To overcome these inefficiencies related to the type conversions introduced by language interoperability, DynQ provides an extension mechanism that can be used to implement language-specific expressions. As discussed in Section 4.2, DynQ relies on two main categories of nodes, namely expression nodes and query operator nodes. Language-specific expressions can be implemented by extending expression nodes with new specializations for types of a certain language.

Considering for example JavaScript dates, language-specific expressions can be implemented to extend comparison nodes by taking care of checking if the objects to be compared are actually JavaScript dates. If so, the comparison can be executed more efficiently by delegating it to the JavaScript engine, an operation that could be inlined by the Graal compiler into DynQ's query operator nodes.

A second usage example of language-specific expressions in DynQ is the implementation of the data-accessor nodes for R data frames. In particular, by leveraging information about R language-specific types, DynQ can access to the schema information within an R data frame before query execution, i.e., at table registration time. Thanks to the language-specific extensions, even if schema information are not required to execute a query with DynQ, it can be still accessed and leveraged to further optimize query execution.

Note that even if we implemented in DynQ typed data-accessor nodes, those typed expressions respects the signature `Object execute(Object)` as shown in Figure 4.2. In this way, also those typed nodes can be composed with all the other expression nodes. As an example, if DynQ needs to evaluate an expression composed of an addition of two fields and the accessor nodes for those fields

are typed nodes, the existing `AddNode` can be used, and it will be specialized to execute an integer addition once the query is JIT compiled.

Moreover, we note that language-specific expressions do not break the high modularity of DynQ, since only expression nodes are extended with such optimizations, whilst adding new query operators, data sources, or features of the query engine (e.g., parallel query execution) would impact only query operator nodes. Moreover, language-specific type expressions are an optional extension, i.e., DynQ can execute queries on objects of a language for which no language-specific expressions are implemented. In this case, depending on the data type of the processed objects, DynQ may have to execute data-conversion operations.

4.5 Fluent API in DynQ

In this section, we focus on data-processing libraries for dynamic languages which allow developers to query heap-allocated objects using a data-frame-like API, i.e., expressing query operators as a chain of method calls. Examples of this syntax are the Spark DataFrame API [2] and LINQ queries when used with the de-sugared method-call syntax [65]. The following is an example of a pipeline built with method chaining, which is a de-sugared version of the LINQ query written with the comprehension syntax in Section 2.1.

```
var evenSquares = xs
    .Where(x => x % 2 == 0)
    .Select(x => x * x);
    .ToArray()
```

Such a chained method-call syntax is used by many existing data-processing libraries. Using this syntax, developers invoke an operator on enumerable objects (also called pipeline builders), passing as parameters the expressions to be evaluated by the operator. The invocation results in a new enumerable object on which another operator can be invoked, forming a chain of method calls. The method chain will result in a materialized result once the developer makes use of a terminal operator, e.g., on the de-sugared LINQ query in the example, `evenSquares.ToArray()` is called to materialize the query result into an array. From now on, we will refer to this syntax as fluent API. As an example of fluent API usage with DynQ, Figure 4.12 shows how the de-sugared LINQ query in the example above can be executed with DynQ.

Note that, in contrast with SQL queries, using a fluent API developers can fragment the definition of a single query. As an example, using a fluent API one

```
var xs = [...];
var evenSquares = DynQ
    .scan(xs)
    .filter(x => x % 2 == 0)
    .map(x => x * x)
    .toArray();
```

Figure 4.12. Example of fluent API usage with DynQ.

can define a function which returns an enumerable object, e.g., the representation of a table scan followed by a predicate, and then call such a function in two different contexts, appending a different terminal operator in each context. This feature greatly improve modularity; indeed, it is often offered by Object-Relational Mapping (ORM) systems [110].

Existing data-processing systems which offer fluent API and that are based on query compilation are implemented similarly to SQL query compilers. In particular, those systems lazily keep track of the operators composing a pipeline as well as their parameters (e.g., UDFs). Once a terminal operator is called, the sequence of operators are considered as a single, standalone query, which is then compiled as a single unit. From now on, we will refer to this approach which triggers compilation for each query execution per-query compilation.

Per-query compilation has the well-known advantage of generating code which is specialized as much as possible for a given query, which means that the generated code is, in principle, the best possible implementation of that query. DynQ offers developers the fluent API leveraging the described per-query compilation approach. However, as further analyzed in the next section and shown in our evaluation, whilst the per-query compilation approach performs very well on analytical workloads, it is suboptimal for high-throughput workloads which perform many queries on small batches of data. In the next section we will present an extension of DynQ to efficiently deal with high-throughput workloads, too.

Chapter 5

Caching Compiled Queries

In the context of query engines based on compilation, a natural solution to the problem of improving the performance on high-throughput workloads is reducing the compilation overhead. Since DynQ is able to execute queries through interpretation before (or instead of) compiling them, compilation overhead is already mitigated. However, for high-throughput workloads, where many queries are executed on small batches of datasets, using the DynQ execution model as discussed before, the application could end up executing all those queries through interpretation.

To improve DynQ's performance on high-throughput workloads, we integrate query-reuse capabilities within the engine. In particular, we present reusable compiled queries, a novel approach to query execution inspired by the code cache implemented in managed runtimes of dynamic languages based on hot-code compilation. With hot-code compilation, the runtime first executes an application through its interpreter. Methods that are invoked often are identified as "hot" and are dynamically compiled to native code. Such an approach has the goal of reaching a stable (or steady) compiler state, i.e., eventually all hot methods which compose the running application are compiled by the JIT of the language runtime. Although the reachability of a stable compiler state is not guaranteed by the runtimes, it is typically achieved for most long-running applications. To reach a stable compiler state, the language runtime must be able to avoid recompiling the same method every time it gets called from a different code location (unless such a method gets inlined). This feature is commonly achieved by leveraging code caches, i.e., map-like data-structures which store the compiled method defined at a given code location.

Considering the context of data-processing libraries, reaching a stable compiler state means that the pipelines are executed in compiled code. However, it

is unlikely that the performance of a single pipeline is as good as the one that could be obtained by compiling the specific pipeline using a per-query compilation approach. Indeed, the compiler might (or might not) decide to inline a certain operator as expected destination of another operator, similarly the compiler might (or might not) decide to inline a whole pipeline in a certain code location, e.g., if detected to be frequently executed.

Table 5.1 shows the benefits and drawbacks of the described approaches, per-query compilation, hot-code compilation, and reusable compiled queries in the context of an application that accepts queries as user input. In particular, per-query compilation can guarantee that all the virtual calls in the implementation of the query operators are de-virtualized through specialization. However, such an approach cannot reach a stable compiler state by design, as every time a query is executed, it triggers its compilation. On the other hand, hot-code compilation is commonly able to reach a stable compiler state, executing the (hot) implementation of the query operators in compiled code. However, method de-virtualization is offered only on a best-effort basis through heuristics. Although it is intuitively impossible to achieve both full de-virtualization and the reachability of a stable state, we argue that by restricting de-virtualization to a subset of calls, it is possible to design an execution model which reduces the number of compiler invocations compared with per-query compilation, but generates more specialized code than hot-code compilation. To implement reusable compiled queries, we first integrate parametricity within the query preparation, such that a single compiled query can be reused multiple times passing different parameters. Then, we leverage the pipeline builders' API to detect similar queries and so that internally we can make use of parametricity, i.e., as an automatic compiler optimization.

In this chapter we first describe parametricity in its simpler form, i.e., prepared statements [78], a well-known feature offered by many database systems. Then, we introduce a parametric extension of a fluent API, a generalization of prepared statements which extends the applicability of parametricity from raw values to expressions through UDFs. Then, we describe reusable compiled queries, a novel approach for implementing a fluent API which does not require developers to make explicit use of parametricity, without suffering from the query compilation overhead for each single query execution as done using the tradition per-query compilation approach. Finally, we present two use case of DynQ. The former shows the benefits of explicit parametricity, while the latter describes a usage scenario where reusable compiled queries are more suitable than explicit parametricity.

Approach	Calls de-virtualization	Reachability of a stable state
per-query compilation	✓ guaranteed (all calls)	✗ never (by design)
hot-code compilation	✗ best-effort (all calls)	✓ most of the applications
reusable compiled queries	✓ guaranteed (subset of calls) ✗ best-effort (remaining calls)	✓ most of the applications

Table 5.1. Benefits and drawbacks of approaches to fluent API compilation: per-query, hot-code, and reusable compiled queries.

5.1 Explicit Parametricity

Prepared statements have been designed to efficiently execute the same query multiple times with differently bound variables. DynQ supports prepared statements which are implemented as instances of `ExecutableNode` that accept parameters. In particular, when a query is prepared, DynQ generates an equivalent AST as discussed in the previous sections. During the AST generation, when DynQ encounters a query variable, i.e., the question mark symbol, it creates an expression node which acts as a placeholder for a value to be bound at query execution time. During the execution of a prepared statement, DynQ binds the placeholders to their values which are retrieved from the local scope of the currently executing query, i.e., its stack frame. Thanks to this approach, once the AST generated from a prepared statement is compiled by the JIT compiler, all subsequent invocations of the prepared statement will execute the same compiled code. Note that, similarly to the case of executing a query without bound parameters, prepared statements may be subject to re-compilation, too. However, since prepared statements are designed to be executed multiple times, re-compilation could take place among different executions. In particular, re-compilation could take place if the types of the prepared statements' parameters change among the different invocations of the same prepared statement, as DynQ would need to generate new machine code specialized for different types.

Figure 5.1 shows an example of a prepared statement with DynQ. The prepared statement in the example is very similar to the example query in Figure 4.3, with the only difference that the expression `x < y` in Figure 4.3 is now `x < ?`. Figure 5.2 shows the AST generated from the prepared statement in Figure 5.1. As expected, the AST is very similar to the one shown in Figure 4.4, the only difference is the node `PlaceHolder$0`, i.e., the placeholder for the prepared statement variable, which replaces the node `ReadMember(y)`. Note that also the com-

```
var data = [{x: 1, y: 2},
            {x: 2, y: 1},
            {x: Date('2000-01-01'), y: Date('2000-01-02')}];

DynQ.registerTable(data, 'T');
var Q = 'SELECT COUNT(*) FROM T WHERE x < ?';
var prepared = DynQ.prepare(Q);
var result1 = prepared(3);
var result2 = prepared(Date('2000-01-02'));
```

Figure 5.1. Example of a DynQ prepared statement on a JavaScript array.

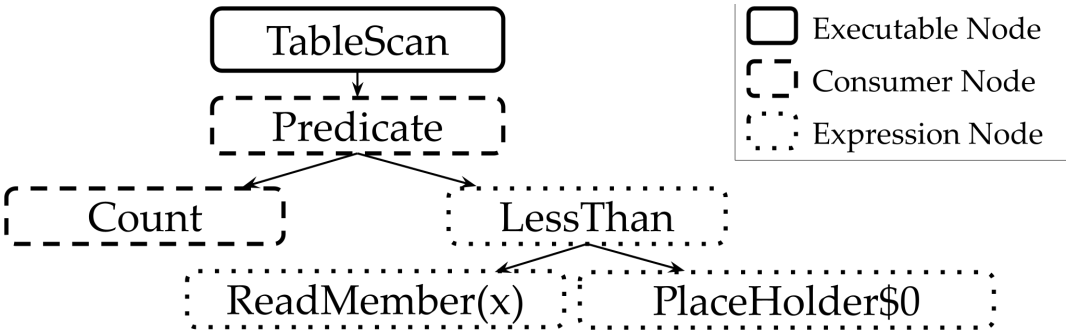

Figure 5.2. AST generated by DynQ for the prepared statement in Figure 5.1.

piled code for the AST depicted in Figure 5.2 is very similar to the one shown in Figure 4.10, with the only difference that the generated function now takes a parameter to be compared with the property `x` of each row.

As we will show in Section 6.3, using prepared statements, DynQ shows performance in line with an equivalent hand-written function which takes the variables of the prepared statement as arguments.

In the implementation of DynQ, we generalize the notion of prepared statement in the context of a fluent API. To this end, we introduced a special marker in our fluent API: `DynQ.par`, as well as a special operator: `prepare`. The parameter marker `DynQ.par` acts as the question mark symbol in prepared statement. However, in contrast with prepared statements, parameters are not limited to placeholder replacements for raw values, since placeholder nodes in the fluent API can represent any UDF. When a UDF is provided as argument to an operator appended into a pipeline-builder, e.g., `.map(x => x*x)`, as with per-query compilation DynQ forces the inlining of the UDF into the query code, de-virtualizing

the call to such UDF. On the other hand, when the marker `DynQ.par` is used to indicate that a parameter will be provided at query execution time, DynQ introduces a virtual call into the generated code pointing to the placeholder location, where DynQ will place the reference to the UDF provided at query execution. Consider again the DynQ fluent API example in Figure 4.12, which selects the squares of the even numbers in a given array. Figure 5.3 shows an example of parametricity defining a similar query which is parametric for the predicate expression. Such a parametric fluent API can be later invoked by passing (as parameter) an arbitrary UDF as predicate expression. Indeed, as the figure shows, the same compiled query can be used to evaluate the squares of even numbers as well as the squares of the numbers which pass any given predicate, e.g., the odd numbers.

As we will show in Section 6.3.3, both prepared statements and parametric fluent API are efficient solutions for query reuse, since the query compilation happens only once and its overhead is mitigated by multiple executions of the same compiled code with different parameters. However, unfortunately, even if parametric fluent API offer a great performance benefit, its usage has many limitations in comparison with a traditional (i.e., non-parametric) fluent API. This is motivated by the fact that both prepared queries and parametric fluent API offer parametricity in an explicit manner. First, all queries must be expressed in the application code, meaning that a system that accepts queries as user input cannot leverage such an approach. Moreover, reusing queries requires developers to carefully refactor each code location in the application which makes use of a fluent API. As an example, in order to leverage the benefit of parametric fluent API, a developer needs to be aware of all the possible code locations within an application which are suitable for being expressed with parametric fluent API, which may not be the case for large applications. Moreover, the process of switching from a common data-processing library with a fluent API to a parametric fluent API version as offered by DynQ may require rewriting a large part of the application. Finally, parametric fluent API cannot be used for cross-library optimizations. In particular, suppose an application makes use of multiple libraries that internally use the same pipeline. To execute that pipeline on the same compiled code, a developer should create a separate module with the definition of the pipeline and refactor those libraries' code such that they all make use of the introduced pipeline in the shared module. In the next sections, we will describe a novel approach for implementing a fluent API which does not require developers to make explicit use of parametricity, without suffering from the query compilation overhead for each single query execution as done using the tradition per-query compilation approach.

```
var xs = [...];
var squaresOf = DynQ
    .prepare()
    .scan(xs)
    .filter(DynQ.par)
    .map(x => x * x)
    .toArray();

var evenSquares = squareOf(x => x % 2 == 0);
var oddSquares = squareOf(x => x % 2 == 1);
```

Figure 5.3. Example of parametric fluent API usage with DynQ.

5.2 Reusable Compiled Queries

In this section we introduce *reusable compiled queries*, a novel approach to query compilation which gets the best of the two above-mentioned approaches, per-query and hot-code compilation. The main design goal of reusable compiled queries is to leverage the DynQ query-compiler to de-virtualize a strict subset of the virtual calls in the implementation of the query operators and to share the same compiled code across multiple similar pipelines. Therefore, also reusable compiled queries are suitable for executing similar pipelines multiple times on small datasets, as well as for executing recursive functions that use the same pipeline with different parameters. However, instead of requiring developers to make explicit use of parametricity, reusable compiled queries internally detect the usage of similar pipelines and leverage parametricity to reuse previously compiled code transparently with respect to the user prospective, i.e., as an automatic compiler optimization.

The subsets of calls that are ensured to be de-virtualized with reusable compiled queries are all the calls to DynQ's Truffle nodes of type `ConsumerNode`, i.e., `consume(row)` and `getResult()`. We avoid forcing inlining of calls to nodes of type `Expression`, leaving the inlining decisions of expression nodes to the underlying JIT compiler (i.e., Graal), as done with all methods during the execution of an application on a VM with hot-code compilation. Consider again the even-squares example query in Figure 4.12, the sequence of operators which composes such a query starts with a source table scan operator, followed by a predicate, a projection, and finally a sink `toArray` operator which materializes the rows into an array. The representation of this query partially specialized by de-virtualizing

the calls to nodes of type `ConsumerNode` is equivalent to the following parametric fluent API.

```
var partiallySpecialized = DynQ
    .prepare()
    .scan(DynQ.par)
    .filter(DynQ.par)
    .map(DynQ.par)
    .toArray();
```

Let's now consider a similar query to the one in the example, i.e., a query which returns the cubes (instead of squares) of the odd numbers (instead of even) in a given array. Since the sequence of operators which composes this query is exactly the same as the even-square query in the example, the partially specialized function shown above can be used for executing both queries, even if their predicate and projection expressions are different. In particular, the odd-cubes pipeline can be executed on the existing compiled code for the function `partiallySpecialized`, passing as parameters `xs`, `x => x % 2 == 1` and `x => x * x * x`.

Reusable compiled queries are an optional feature of our fluent API, which developers can enable globally as well as for a single query. We implemented the caching strategy behind reusable compiled queries within the pipeline builders, leveraging the parametric fluent API described in Section 5.1. In particular, the reusable compiled queries are stored with a tree shape in a memory location which is shared among the whole application. The pipeline tree is composed of two kinds of nodes, intermediate nodes and leaves nodes. Each leaf contains a prepared query generated with a parametric fluent API, whilst each intermediate node represents a query operator. Thus, a path from the root to a leaf represents a sequence of operators, i.e., a pipeline, and such a leaf contains a reference to the compiled representation of that pipeline, i.e., a DynQ executable node. Since there must be a single root in a tree and there can be multiple scan implementation, the root of pipelines tree is an empty operator; all scan operators are children of the empty root.

Reusable compiled queries are transparently created by DynQ through the pipeline builder instances created when developers make use of a fluent syntax. Figure 5.4 shows the internal (simplified) Java implementation of the pipeline builders, and Figure 5.5 shows the pipeline nodes implementation used by the builders. As the figure shows, each pipeline-builder object contains two fields, a reference to a (shared) node in the pipeline tree, and an array of actual parameters. Each node in the pipeline tree contains three fields, a reference to

a (partially) prepared query using a parametric fluent API, a map which stores the children nodes by operator, and a reference to an executable node generated by the (fully) prepared query, which is non-null only for leaf nodes. Note that the empty root of the pipeline tree is created with an empty map and a reference to a prepared query with parametric fluent API without any operator, i.e., `root.query = DynQ.prepare()`.

When a developer makes use of a fluent API with reusable compiled queries, DynQ internally creates a new pipeline builder composed of an empty array as actual parameters and a reference to the (shared) root of the pipeline tree. Then, every time an operator Op with parameters $(p_1, ..., p_n)$ is appended on a pipeline builder P_0 through method call, the method $P_0.appendOperator$ is called, and a new pipeline-builder instance P_1 is created (line 8 in Figure 5.4). The actual parameters of P_1 are defined as $P_0.actualParameters$ ++ $[p_1, ..., p_n]$, where ++ denotes the array concatenation (line 9 in Figure 5.4). The reference to a the pipelines-tree node in P_1 (i.e., $P_1.sharedNode$) will be evaluated as follows. If the node $P_0.sharedNode$ has already a child node for the operator Op (say $node'$), then we define $P_1.sharedNode = node'$ (line 18 in Figure 5.4). Otherwise, a new (shared) node $node''$ is created for the operator O as a child of $P_0.sharedNode$ in the pipeline tree, and $P_1.sharedNode$ will be assigned to $node''$ (lines 20-32 in Figure 5.4). Finally, if the operator Op is a terminal operator, then $P_1.sharedNode$ is a leaf in the pipeline tree. If such a leaf was freshly created, DynQ creates the `ExecutableNode` through the parametric fluent API instance in the tree node and cache it in the tree node itself. Otherwise DynQ reuses the already generated (i.e., cached) `ExecutableNode`. Note that the generated AST does not contain any expression node but parameters, since all actual parameters provided by the developer are internally stored in the pipeline-builder instances and replaced with `DynQ.par` within the generated executable node, making that node reusable for any other query composed of the same sequence of operator. Once the `ExecutableNode` has been retrieved (either cached or freshly created) it is automatically invoked by passing as parameters the array $P_1.actualParameters$ (line 18 in Figure 5.4). It is important to note that reusable compiled queries do not prevent additional speculative compiler optimizations, e.g., the compiler could decide to speculatively inline a UDF in a query as in hot-code compilation.

Note that reusable compiled queries share a similar design goal with the code cache implemented in language runtimes, i.e., avoiding recompiling the same code location multiple times. There are however important differences between reusable compiled queries and general-purpose code caches. In particular, a code-cache can store a compiled method, it cannot automatically partially

```
class PipelineBuilder {

    Object[] actualParameters;
    PipelineNode sharedNode;

    Object appendOperator(Operator op, Object[] params) {

        PipelineBuilder next = new PipelineBuilder();
        next.actualParameters = arrayConcat(
            this.actualParameters, params);

        next.sharedNode = this.sharedNode.getOrCreate(op);
        if(next.executable != null) {
            // terminal operator:
            // invoke the executable node
            // and return the result
            return next.executable.execute(
                next.actualParameters);
        } else {
            // intermediate operator:
            // return the new enumerable object
            return next;
        }
    }

}
```

Figure 5.4. Java-like pseudocode of PipelineBuilder objects for reusable compiled queries.

```
class PipelineNode {
    static class ParametricFluent {
        // Implementation omitted for brevity
        ParametricFluentAPI append(
            Operator op, Object[] params) { ... }

        ExecutableNode toExecutable() { ... }
    }

    ParametricFluent query;
    Map<Operator, PipelineNode> children;

    // note: non-null only for leaf nodes
    ExecutableNode executable;

    PipelineNode getOrCreate(Operator op) {
        if(children.contains(op)) {
            return children.get(op);
        }
        int n = op.parametersCount;
        Object[] params = new Object[n];
        for(int i=0; i<n; i++) {
            params[i] = DynQ.par;
        }

        PipelineNode next = new PipelineNode();
        next.query = this.query.append(op,params);
        if(op.isTerminal) {
            next.executable = next.query.toExecutable();
        }
        children.put(op, next);
        return next;
    }
}
```

Figure 5.5. Java-like pseudocode of PipelineNode objects used by the PipelineBuilder implementation in Figure 5.4.

specialize such a method for subsequent similar reuse, as in the case of reusable compiled queries. Moreover, the lookup in the code cache is more expensive than the one in the pipeline tree, as each lookup is performed on a single global map on the code cache, whilst in the case of the pipeline tree, each lookup is local at the level of a specific operator. In particular, the cost of a single lookup in the pipeline tree is constant, i.e., a lookup only checks whether a specific attribute of an object is `null`. Reusable compiled queries can be seen as an optimized data-processing-specific code cache for fluent API which is able to detect similar queries and reuse their compiled representation.

Thanks to the normal (i.e., non-parametric) fluent syntax offered by reusable compiled queries, a developer can switch from using a typical data-processing library implemented in a dynamic language to DynQ with minimal effort. In particular, there is no need to take care of rewriting the pipelines with explicit parametricity leveraging the parametric fluent API. Moreover, as we will show in Section 6.3.3, thanks to reusable compiled queries, DynQ outperforms Lodash [66], a popular data-processing library for JavaScript, making DynQ an attractive drop-in replacement for existing data-processing libraries.

We note that reusable compiled queries may reduce peak performance of long-running queries when compared to per-query compilation. This is expected, because per-query compilation guarantees that all calls in the query-operator implementations are de-virtualized at compilation time, which is not the case for reusable compiled queries. However, it is important to note that, when reusable compiled queries are enabled, the virtual calls that are not de-virtualized by DynQ are still candidates to be de-virtualized by the underlying VM on a best-effort basis, which in many cases is effectively applied. As an example, if a hot code location makes use of the DynQ's fluent API and the calls to DynQ are inlined within that code location, the parameters passed to the fluent API (e.g., the UDF for a `filter` operation) can be de-virtualized by the VM once the generated code for DynQ's fluent API is inlined and optimized in the context of the caller.

5.3 DynQ Use Cases

Here, we present two realistic use cases of DynQ. As mentioned in Chapter 1, JavaScript and Node.JS are widely used to implement data-intensive server-side applications. As realistic use cases of DynQ, we show how it can be exploited for (re-)implementing two existing npm [77] modules, highlighting the benefits of explicit parametricity and of reusable compiled queries.

Explicit parametricity The first use case is the npm module cities [17], which exposes a dataset of locations and offers an API for selecting and filtering elements. Since in the cities module all the API is implemented as functions which take simple, raw values as parameters (e.g., the prefix of the locations to be selected), we implemented the API in DynQ leveraging the prepared statements introduced in Section 5.1. As an example, below is the DynQ implementation of the `findByState` API, which finds the first location that matches a given state name.

```
var locations = { ... }
DynQ.registerTable(locations, 'locations');
Q = 'SELECT * FROM locations WHERE state=?';
findByState = DynQ.prepare(Q);
```

The cities npm module could also be implemented using an RDBMS to query the set of locations. However, if the database ran in a separate process (or, even worse, on a different machine), each query execution would suffer from inter-process communication overhead, as well as from serialization and deserialization overheads [88]. If the database was embedded in the language (i.e., both the database engine and the language runtime execute in the same process), each query result set would have to be converted to heap-allocated objects before it could be accessed by JavaScript code. That is, also such a more efficient usage of a database would introduce data-conversion overheads.

Instead of relying on databases, the developers of the cities module opted for using Lodash [66], a popular JavaScript data-processing library, for querying the locations, a solution which avoids any data conversions. Also the DynQ implementation of the API in the cities module does not need any data conversions. Moreover, as we will show in Section 6.3.2, DynQ is faster than Lodash for the cities module. Besides performance, the DynQ implementation of the API defined in the cities module is more concise (and therefore easier maintainable) than the original implementation using Lodash.

Reusable compiled queries As a second realistic use case, we consider the npm module json-server [49]. This module offers a REST [69] server which loads JSON files into memory (i.e., as JavaScript arrays of objects) and exposes an HTTP API to query those array-based tables.

We consider json-server a good candidate for the reusable-compiled-queries approach implemented in DynQ. Indeed, json-server is implemented internally using Lodash, and it creates data-processing pipelines in a dynamic way. In particular, a simplified implementation of the json-server main entry point is de-

picted in Figure 5.6. As the figure shows, each pipeline is built dynamically within a for loop, conditionally appending query operators depending on the given input parameters provided in the HTTP request. In such a scenario, it would be challenging for a developer to use explicit parametricity to speed up query execution. The implementation would need to detect similar pipelines by inspecting the HTTP parameters and eventually selecting one of the previously compiled queries for each incoming request, or creating a new query if there is no previously compiled one for the given set of parameters.

The resulting implementation would be an ad-hoc cache of compiled queries manually built by the application developer, resorting to explicit parametricity, which is very similar to how reusable compiled queries are internally implemented. On the other hand, leveraging reusable compiled queries allows the developers to leave to DynQ the complexity of managing the cache of compiled queries and to keep the implementation much simpler, exactly as shown in Figure 5.6, only replacing `lodash.chain` with `DynQ.scan`.

```
function entryPoint(table, urlParameters) {
    pipeline = lodash.chain(table)
    for(param of urlParameters) {
        pipeline = appendOperatorFor(param, pipeline)
    }
    return pipeline.toArray()
}
function appendOperatorFor(param, pipeline) {
    if(param.key == 'lt')
        return pipeline.filter(x => x < param.value)
    // remaining operators omitted for brevity
}
```

Figure 5.6. Simplified main entry point of the json-server npm module.

Chapter 6

DynQ Evaluation

In this section we evaluate the performance of DynQ. First, we describe our evaluation plan(Section 6.1), explaining the setup and the motivation for each experiment. Then, we evaluate DynQ with two dynamic languages, R (Section 6.2) and JavaScript (Section 6.3). We evaluate DynQ on existing, established workloads designed for both databases and programming languages. Moreover, we also evaluate DynQ in a realistic scenario on existing' code bases, by recasting an existing server-side data-processing application to make use of DynQ.

As database workloads, we evaluate DynQ using the TPC-H benchmark [111] queries and a micro-benchmark composed of a set of queries based on the dataset of the TPC-H benchmark. Those queries, listed in Table 6.1, have been presented in the context of a *stream-fusion engine* [97], and they belong to the following categories:

- Queries consisting of selection and aggregation (without group by), leading to a single row (i.e., queries 1, 2, 3).
- Queries consisting of selection, projection, which return a list of rows (i.e., query 4), with also a limit operator (i.e., query 6) and with both sort and limit (i.e., query 5).
- A query consisting of selection and join, followed by an aggregation operator resulting in a single row (i.e., query 7).

From now on, we refer to the *i*-th query in TPC-H as Q*i*, and to the *j*-th query in the micro-benchmark as MQ*j*.

We run all our experiments on an 8-core Intel i9-10980XE (@3.0 GHz) with 256 GB of RAM. The operating system is a 64-bit Ubuntu 20.04 and the language runtime is GraalVM Community Edition 21.3.0, i.e., the latest LTS release at the

Table 6.1. Micro-benchmark queries from stream-fusion engine [97].

MQ1	`SELECT COUNT(*)`
	`FROM lineitem`
	`WHERE l_shipdate >= DATE '1995-12-01'`

MQ2	`SELECT SUM(l_discount * l_extendedprice)`
	`FROM lineitem`
	`WHERE l_shipdate >= DATE '1995-12-01'`

MQ3	`SELECT SUM(l_discount * l_extendedprice)`
	`FROM lineitem`
	`WHERE l_shipdate >= DATE '1995-12-01'`
	`AND l_shipdate < DATE '1997-01-01'`

MQ4	`SELECT l_discount * l_extendedprice`
	`FROM lineitem`
	`WHERE l_shipdate >= DATE '1995-12-01'`

MQ5	`SELECT l_extendedprice`
	`FROM lineitem`
	`WHERE l_shipdate >= DATE '1995-12-01'`
	`ORDER BY l_orderkey`
	`LIMIT 1000`

MQ6	`SELECT l_discount * l_extendedprice`
	`FROM lineitem`
	`WHERE l_shipdate >= DATE '1995-12-01'`
	`LIMIT 1000`

MQ7	`SELECT SUM(o_totalprice)`
	`FROM lineitem, orders`
	`WHERE l_orderkey = o_orderkey`
	`AND l_shipdate >= DATE '1995-12-01'`
	`AND o_orderdate >= DATE '1995-12-01'`

time of writing. Unless otherwise specified, for all experiments the reported execution times include the query preparation time, i.e., the Truffle nodes generation obtained by traversing the query plan generated by Calcite and the actual query execution time. Note that we do not measure the time spent for query parsing and planning done by Calcite since it is not an optimized component of our system and, on some queries, planning is currently rather slow on Calcite, a performance issue which can be solved with more engineering effort. However, it is important to note that measured time takes into account the generation of our physical plan representation (i.e., Truffle nodes), and also their JIT compilation, which happens during query execution. Unless otherwise indicated, all the figures presented in this section are bar plots that show the query execution time for each implementation. The numbers on top of the bars represent the speedup (factors) achieved by DynQ. Speedup factors below 1 indicate that DynQ is slower.

6.1 Evaluation Plan

On R, we evaluate DynQ against the `data.table` API, DuckDB [89], and MonetDB [45]. In this setting, we import the TPC-H tables into R data frames. Since TPC-H is based on a strict (relational) schema, and the data is imported into R data frames, which is a typed data structure, the evaluation on R does not highlight the DynQ peculiarity of efficiently accessing data with unknown schema. Indeed, for all the experiments in this setting, DynQ uses the schema information from the data frames. However, DynQ currently uses the schema only for the data-access operations, all the query operators nodes as well as the other expression nodes share the same implementations as in the case of unknown schema, as described in Section 4.4. The main goal of this evaluation is to show that on relational database workloads the flexibility of DynQ in accessing data formats which are not directly managed by the query engine does not impair performance, in contrast to other data-processing systems.

On JavaScript, we evaluate DynQ in very different settings. First, we evaluate DynQ against AfterBurner [28] using AfterBurner's memory layout, i.e., a columnar layout composed of typed JavaScript arrays. In this setting we use TPC-H and the microbenchmark queries as workloads. Since AfterBurner is a relational database, also this setting uses a strict schema, and similarly to the evaluation on R the goal of this evaluation is to show that query execution performance with DynQ on relational data is in line with a query engine which reads data using its own memory layout.

Then, we evaluate DynQ on datasets stored as JavaScript object arrays. For all the experiments in this setting, no schema information is provided to DynQ. Here, we first evaluate DynQ against Lodash [66] and hand-written implementations using the microbenchmarks. Then, we evaluate DynQ on existing code bases (the npm module *cities* [17]), leveraging prepared statements (Section 6.3.2) to implement a web-service backend module to search locations based on user input. Those experiments highlight the ability of DynQ to efficiently process dynamic objects with unknown schema. Finally, we evaluate reusable compiled queries (Section 6.3.3) on a JavaScript implementation of two relevant benchmarks that use a fluent API for data processing that we recasted from the Renaissance [85] benchmark suite, which was originally implemented in Java. One of these two latter experiments also show the ability of DynQ to handle efficient query execution on polymorphic types, since the engine needs to deal with mixed types of input arrays.

6.2 R Benchmarks

In this section we evaluate DynQ with the R programming language. Here, we use the dataset from the TPC-H benchmark generated with the original dbgen tool [111] loaded into an R data frame. Since, like DynQ, DuckDB [89] allows executing SQL queries directly on R data frames, we evaluate DynQ on the TPC-H benchmark queries and the micro-benchmark queries against DuckDB, on a dataset of scale factor 10; the dataset size is 10GB in a text format. In particular, we use DuckDB (version 0.3.0), executed on GnuR [86] (version 3.6.3). Since the measured execution time with DynQ does not take into account query planning time, we slightly modified the DuckDB R plugin so that queries can be planned and executed in two different steps, so that the measured execution time on DuckDB does not take into account query planning as well. DuckDB provides two ways for executing queries on R data frames, i.e., directly on the data-frame data structure, and in a managed table, which is much more efficient but requires an ingestion phase. We refer to the former setting as DuckDB(df), and to the latter one as DuckDB(preload). Note that, by comparing DynQ against DuckDB, the fair comparison is with DuckDB(df), since the data is accessed directly on R data frames, as in DynQ. Moreover, in evaluating DuckDB(preload), we do not measure the time spent in the ingestion phase. In this evaluation, we measure the median of 10 executions.

Due to the different query planners and implementation choices in DynQ and DuckDB (DuckDB is vectorized and interpreted whilst DynQ is tuple-at-a-time

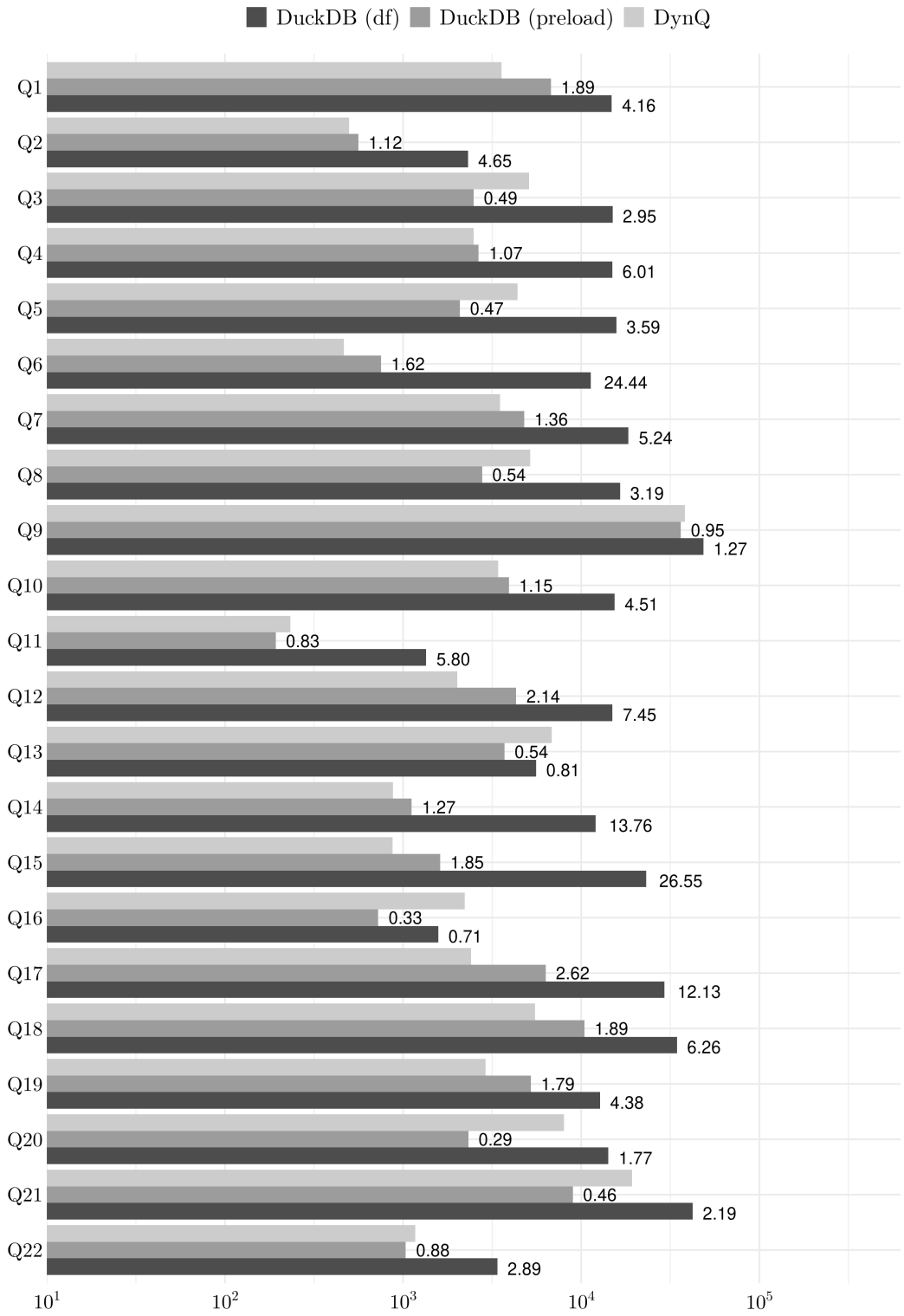

Figure 6.1. R TPC-H benchmark (SF-10).

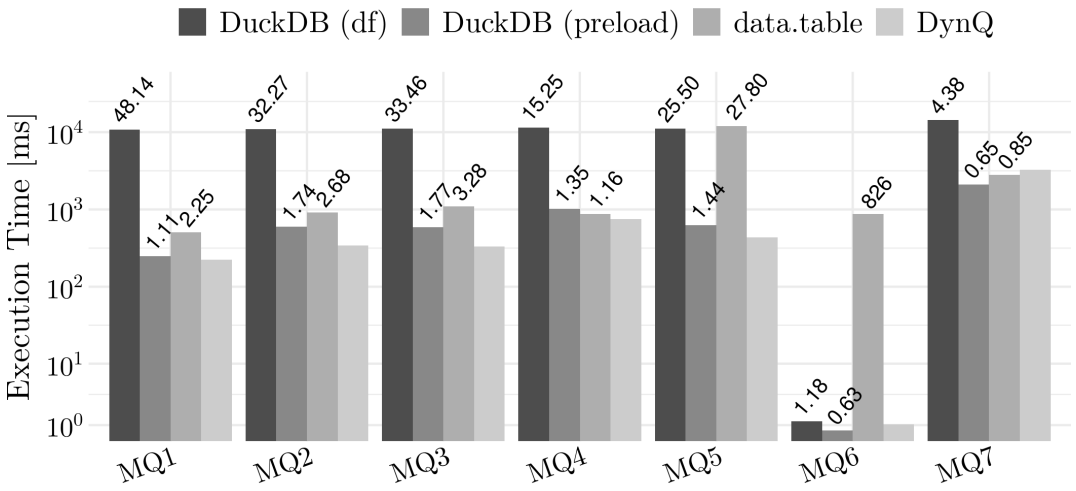

Figure 6.2. R micro-benchmark (SF-10).

and JIT compiled), the goal of this performance evaluation is not to compare two very different systems, but rather to demonstrate that DynQ achieves performance competitive with an established, state-of-the-art data-processing system. We consider the micro-benchmark queries important in our evaluation, since, due to their simplicity, the query plans are the same in DynQ and in other systems. Moreover, since the micro-benchmark queries are rather simple, they stress data-access operations, showing that the extensibility of DynQ in accessing data in different formats does not impair query execution performance, which we consider a great achievement.

Micro-benchmarks. Due to the simplicity of the queries in the micro-benchmarks listed in Table 6.1, we manually implement them using the `data.table` API, which is arguably the most efficient library for processing R data frames. The benchmark results are depicted in Figure 6.2. As the figure shows, DynQ is slower than the `data.table` API only on MQ7 and outperforms it on all other queries by speedup factors ranging from 1.16x (MQ4) to 27.8x (MQ5). The speedup on MQ6 against `data.table` (i.e., 826x) is because DynQ chains query operators and stops the computation once it finds the first 1000 elements that satisfy the predicate (i.e., the limit operator). On MQ6, DynQ performs comparably with DuckDB, with speedup factors of about 1.18x against DuckDB(df) and 0.63x against DuckDB(preload), with a query execution time of about 1ms, showing the effectiveness of our exception-based approach for implementing early exits for the LIMIT operator. We consider such a low query execution time a great achieve-

ment for DynQ, since the existing query engines based on compilation commonly suffer from a latency overhead due to query compilation. DynQ outperforms DuckDB(df) in all other queries as well, with speedup factors ranging from 4.38x (MQ7) to 48.14x (MQ1). Those speedups against DuckDB(df) are motivated by the fact that the micro-benchmark queries are simple and mostly dominated by table scans. DuckDB(df) requires table-scan operations to convert data on-the-fly from data frames into the DuckDB physical data representation, which introduces high overhead. On the other hand, DynQ can execute queries on R data frames in-situ, i.e., without any conversion. Indeed, DynQ performance is closer to DuckDB(preload), which significantly outperforms DuckDB(df), showing that the great flexibility of DynQ in accessing data in different formats does not impair performance. In particular, DynQ is slower than DuckDB(preload) only on queries MQ6 (factor 0.63x) and MQ7 (factor 0.65x), and outperforms DuckDB(preload) on all other queries, with speedup factors ranging from 1.11x (MQ1) to 1.77x (MQ3).

TPC-H Benchmark. Here, we evaluate DynQ using the TPC-H benchmark. Like in our previous experiment, we compare DynQ against DuckDB executing queries directly on the data frame, i.e., DuckDB(df) and with data loaded into a managed memory space, i.e., DuckDB(preload). The benchmark results are depicted in Figure 6.1.

As the figure shows, DynQ is slower than DuckDB(df) only on Q13 (factor 0.81x) and Q16 (factor 0.71x), in all other queries DynQ outperforms DuckDB(df), with speedup factors ranging from 1.27x (Q9) to 26.55x (Q15). In comparison with DuckDB(preload), DynQ is faster on 12 queries (i.e., Q1, Q2, Q4, Q6, Q7, Q10, Q12, Q14, Q15, Q17, Q18, Q19).

Latency Benchmarks. As discussed in Section 4.2, even if DynQ is an engine based on query compilation, it is able to start executing a query before compiling it, by executing the Truffle nodes which represent the query in the interpreter. This feature is crucial for obtaining high throughput when executing queries on small datasets. Here, we evaluate the throughput of DynQ against DuckDB. Since DuckDB is based on interpretation and vectorization, it does not spend any time on code generation and query compilation. On small datasets, this approach is commonly faster than compiling queries, since the compilation overhead may not be paid off.

For this evaluation, we consider an experiment similar to the one performed in the context of Umbra [76]. Such an experiment [52] evaluates the throughput

(by calculating the geometric mean of queries per second for all TPC-H queries) over different scale factors. In our experiment we evaluate the throughput over scale factors 0.001, 0.01, 0.1, 1, and 10, first on the micro-benchmark queries and then on the TPC-H queries.

The benchmark results are depicted in Figure 6.4 for the micro-benchmark and in Figure 6.3 for TPC-H. As the figures show, for both the micro-benchmark and TPC-H, DynQ outperforms DuckDB(df) on all evaluated scale factors. In particular, on the micro-benchmark DynQ outperform DuckDB(df) of factors 9.17x (SF 0.001), 5.94x (SF 0.01), 6.51x (SF 0.1), 11.02x (SF 1) and 14.29x (SF 10). On TPC-H, DynQ outperform DuckDB(df) of factors 5.59x (SF 0.001), 3.56x (SF 0.01), 2.65x (SF 0.1), 4.31x (SF 1) and 4.4x (SF 10).

In comparison with DuckDB(preload), the evaluation shows interesting trends. On the smallest scale factors (SF 0.001 and 0.01), DynQ fully executes all queries in the interpreter and it never triggers compilation. On scale factor 0.001, the DynQ throughput differs from DuckDB(preload) by a factor of 4.08x on the micro-benchmark, and of 2.7x on TPC-H. On scale factor 0.01, the DynQ throughput is in line with DuckDB(preload), in particular, DynQ shows a throughput factor improvement of 1.06x on the micro-benchmark, and of 0.95x on TPC-H. On scale factors 0.1, DynQ starts compiling parts of the queries; however, since the datasets are still small, most of the query execution is still in the interpreter. On such scale factor, the DynQ throughput is smaller than the one of DuckDB(preload), by factors 0.58x on the micro-benchmark, and of 0.56x on TPC-H. Then, on scale factor 1, in DynQ query compilation is paid off on the micro-benchmark, reaching a throughput in line with the one of DuckDB(preload), i.e., 0.96x factor. This is not the case for TPC-H, where the throughput of DynQ is factor 0.84x compared with DuckDB(preload). The reason is that the TPC-H queries are much more complex than the micro-benchmark queries, leading to longer query compilation times. Finally, on scale factor 10, DynQ outperforms DuckDB(preload) on the micro-benchmark by a factor of 1.15x, and becomes comparable with DuckDB(preload) on the TPC-H queries, by a factor of 0.98x.

Our evaluation on the query latency shows that JIT compilation in DynQ is not a source of performance concerns, differently from most existing query engines based on compilation.

Comparison with Native DBMS. In this section we evaluate DynQ against MonetDB [45], a modern, interpreter-based, RDBMS featuring high-performance vectorized execution. Although we do not consider MonetDB a direct competitor to DynQ, this evaluation should be considered an indication of how DynQ

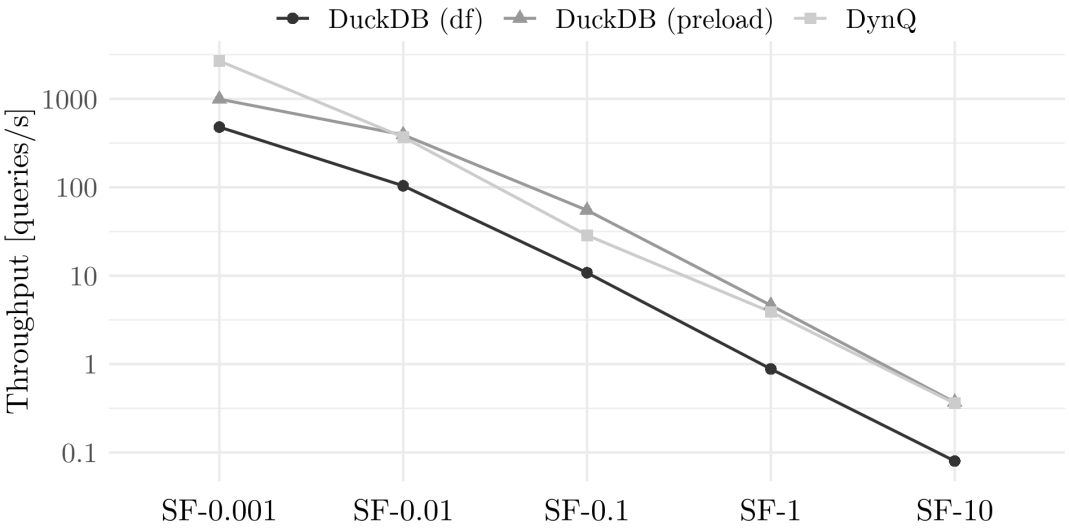

Figure 6.3. Geometric mean of queries/s (TPC-H).

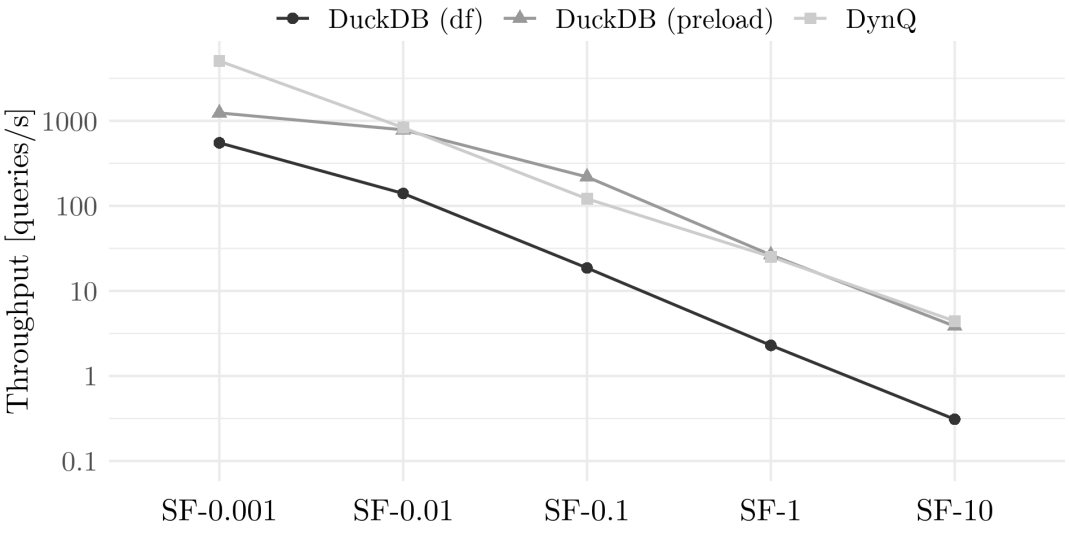

Figure 6.4. Geometric mean of queries/s (micro-benchmark).

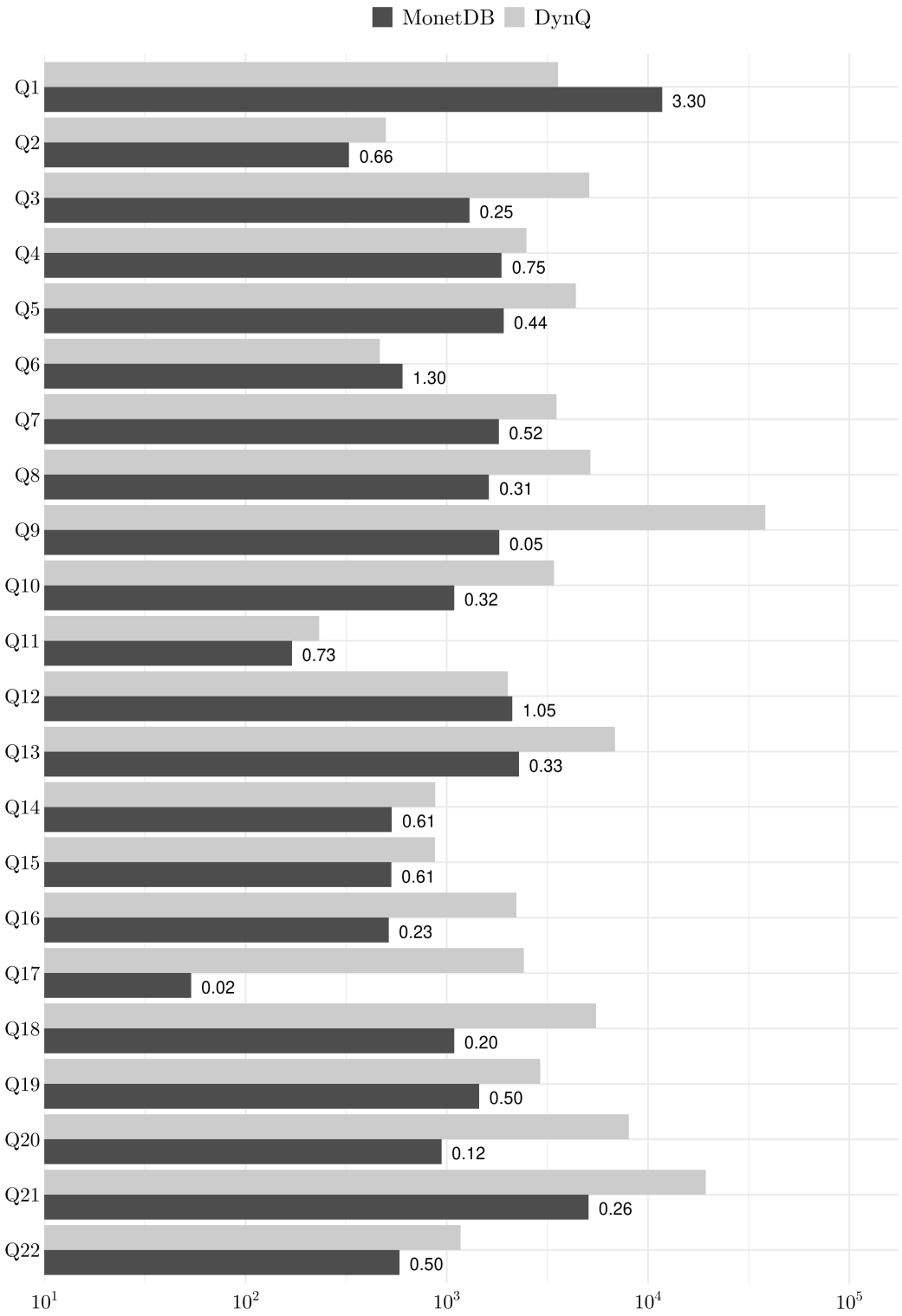

Figure 6.5. TPC-H benchmark against MonetDB (SF-10).

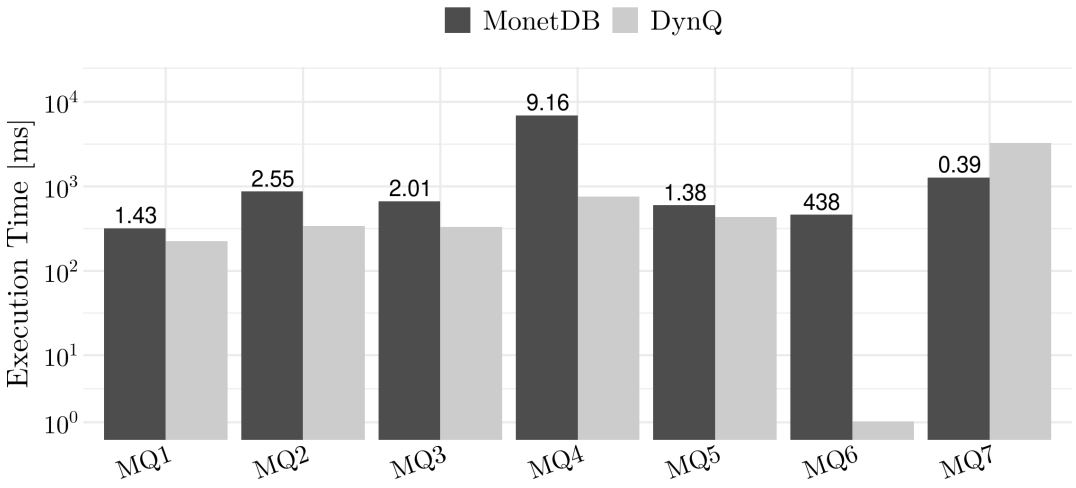

Figure 6.6. Micro-benchmark against MonetDB (SF-10).

performs in comparison with a native RDBMS. For this evaluation, we use MonetDB Database Server Toolkit 11.43.9 (Jan2022-SP1), executing the queries with mclient. We measure the end-to-end query execution time, taking into account the cost of inter-process communication for sending result sets from the server to the client process. For fairness, we configure MonetDB for executing in a single-thread, since we have not yet implemented parallel query execution in DynQ. For this experiment, we evaluate DynQ on R data frames, using a scale factor of 10 for both the micro-benchmark and TPC-H; we present the median of 10 executions.

The benchmark results are depicted in Figure 6.5 for TPC-H and in Figure 6.6 for the micro-benchmark. As the figures show, MonetDB outperforms DynQ in all queries containing the join operator, i.e., in MQ7 and in all TPC-H queries but Q1 and Q6, with the only exception of Q12, where MonetDB and DynQ show very similar execution times. All remaining queries are rather simple and mostly dominated by table scans. For those queries, DynQ is faster than MonetDB; in particular DynQ outperforms MonetDB by a speedup factor of 3.3x on Q1 and 1.3x on Q6. Concerning the remaining micro-benchmark queries, on MQ6 DynQ shows a speedup of 438x; this is because MonetDB (like the data.table R package) does not stop the query execution once the first 1000 elements (i.e., the limit operator) have been found. On MQ4, DynQ outperforms MonetDB by a speedup factor of 9.16x, because MQ4 returns a large result set that MonetDB needs to serialize and transfer to the client process, whereas DynQ (being an embedded query engine) does not incur such an overhead. Finally, on MQ1, MQ2, and MQ3,

MQ5 DynQ outperforms MonetDB by speedup factors of 1.43x, 2.55x, 2.01x, and 1.38x.

6.3 JavaScript Benchmarks

Here, we evaluate DynQ using the JavaScript programming language. For this evaluation, we first compare DynQ against AfterBurner [28], which is an in-memory database entirely written in JavaScript, on both the micro-benchmark and on TPC-H. Then, we evaluate DynQ querying data loaded into a JavaScript array of objects, like in the example of Figure 4.3. In this setting, we evaluate DynQ on the micro-benchmark against hand-written implementations in JavaScript and implementations that rely on Lodash [66], which is arguably the most efficient and popular data-processing library for JavaScript. Finally, we evaluate DynQ on existing code bases, comparing the performance of a JavaScript library against equivalent implementations using DynQ.

6.3.1 Evaluation on AfterBurner

For evaluating DynQ against AfterBurner [28], we implemented a specific DynQ provider for the memory layout implemented in AfterBurner, i.e., a columnar layout composed of JavaScript typed arrays. The implementation of such a specific data-source provider required only about 1000 lines of code, which shows the great extendibility of DynQ. In this setting we evaluate AfterBurner both on GraalVM and on V8 [112] (Node.JS version 14.17.6). All our experiments on AfterBurner are executed using only scale factor 1; we cannot evaluate AfterBurner on bigger datasets due to a limitation in the Node.js file parser used in AfterBurner, which cannot parse files exceeding 2GB. In this setting, we measure the median of 20 executions.

Micro-benchmarks. Due to the simplicity of the queries in the micro-benchmark listed in Table 6.1, we manually implemented them using the AfterBurner API, which is a fluent API inspired by Squel.js [102]. The benchmark results are depicted in Figure 6.8. As the figure shows, even if AfterBurner is based on query compilation, it does not optimize the early exit for the limit operator. Thus, for MQ6, DynQ outperforms AfterBurner by a speedup factor of 145x on V8, and 826x on GraalVM. DynQ outperforms AfterBurner running on GraalVM for all other queries, too, ranging from a speedup factor of 2.56x (MQ4) to 12.33x (MQ5). When executing AfterBurner on V8, AfterBurner is faster than DynQ

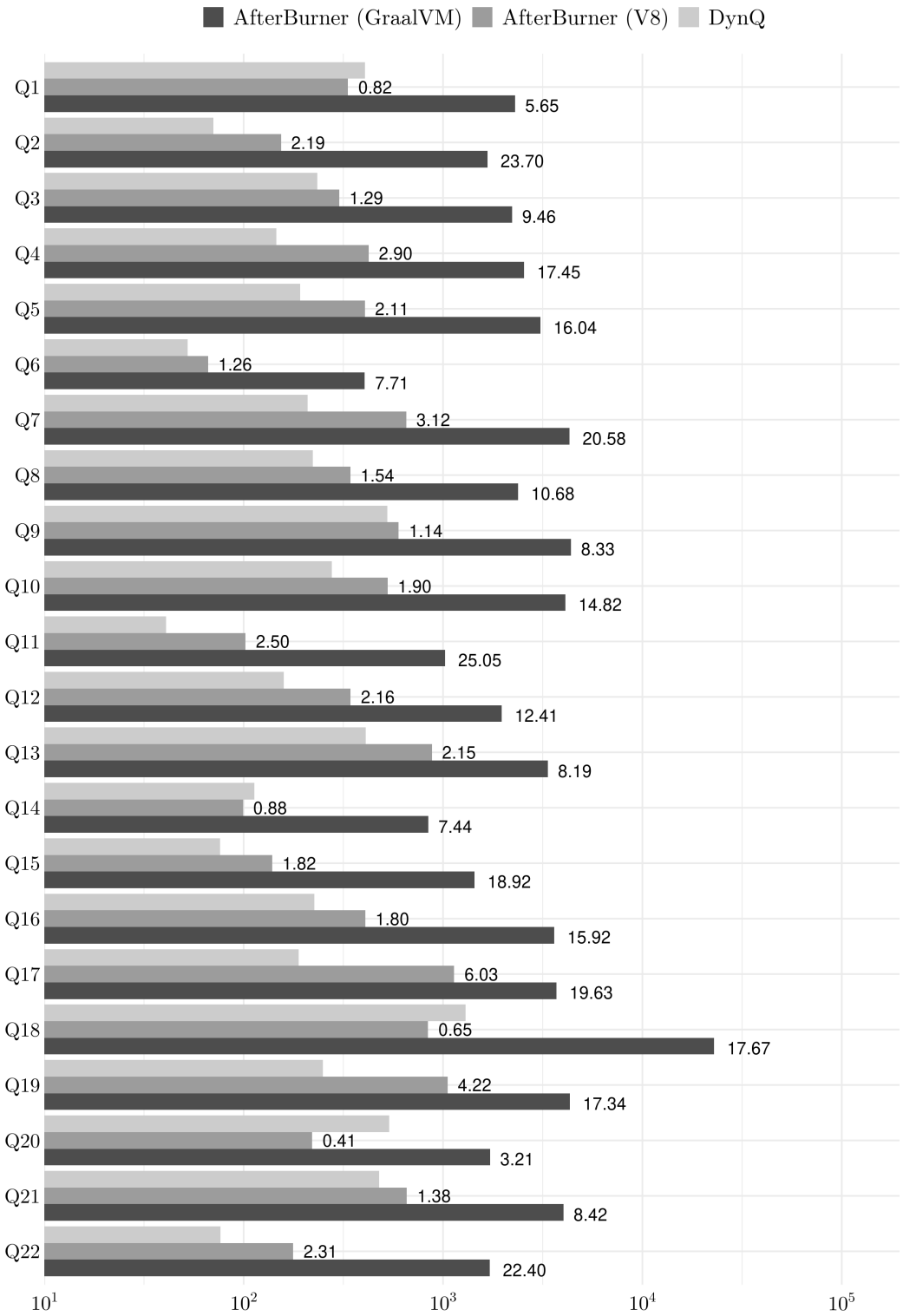

Figure 6.7. JS TPC-H benchmark on AfterBurner (SF-1).

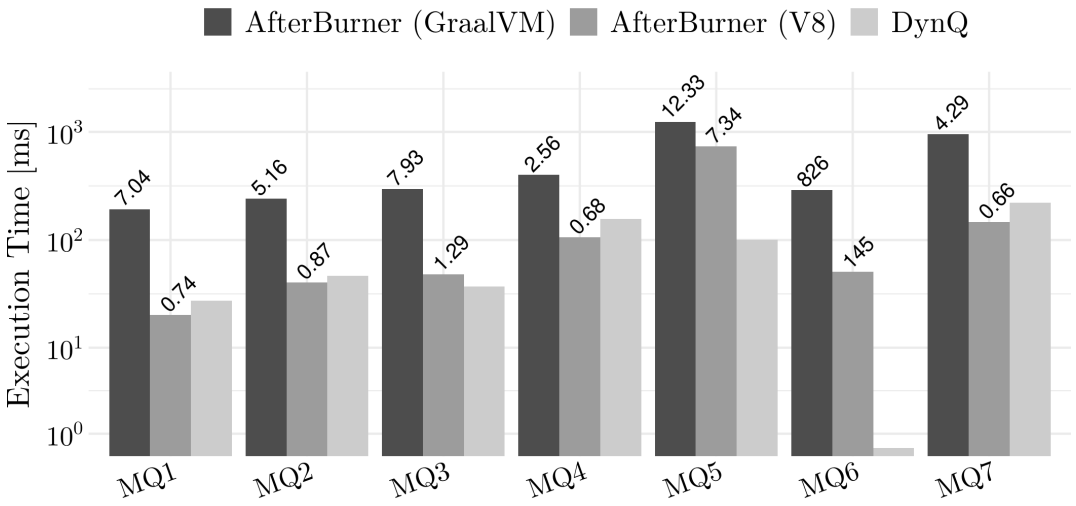

Figure 6.8. JS micro-benchmark on AfterBurner (SF-1).

on MQ1, MQ2, MQ4 and MQ7; the reason is that V8's compiler is faster than GraalVM on these queries, so the benefit of compilation is almost immediate.

TPC-H Benchmark. We evaluate DynQ against AfterBurner on TPC-H using the original AfterBurner benchmark [1]. Since AfterBurner uses a fluent API, there is no query parsing and planning phase, and the query plan is made explicit by the API usage. For fairness, we manually fine-tuned the queries in our evaluation such that Calcite generates the same query plans used by AfterBurner. The benchmark results are depicted in Figure 6.7. As the figure shows, DynQ outperforms AfterBurner executed on GraalVM on all queries, with speedup factors ranging from 3.21x (Q20) to 25.05x (Q11). When executing AfterBurner on V8, DynQ is slower on queries Q1 (0.82x), Q14 (0.88x), Q18 (0.65x) and Q20 (0.41x). On all remaining queries, DynQ outperforms AfterBurner on V8 with speedup factors ranging from 1.26x (Q6) to 6.03x (Q17), since AfterBurner materializes more intermediate results than DynQ.

6.3.2 Evaluation on Object Arrays

Here, we evaluate DynQ using JavaScript object arrays as datasets. First, we evaluate DynQ on the micro-benchmark against equivalent hand-written implementations. Then, we evaluate DynQ on an existing code base, by comparing the original implementation of an npm [77] module with an equivalent one based on DynQ. Here, we measure query execution time at peak performance.

Micro-benchmarks. Similarly to the evaluation on R, we manually implemented the micro-benchmark queries in the JavaScript language. In this setting, we evaluate the micro-benchmark queries against hand-written implementations and implementations that use Lodash. Since Lodash does not offer an API for the join operator, we do not evaluate MQ7 using Lodash. The scale factor used for our JavaScript evaluation is 1 (whereas we used a scale factor of 10 for the R evaluation). This is motivated by the fact that querying R data frames is more efficient than JavaScript object arrays, since R data frames are internally implemented using a columnar data format composed of typed arrays, whereas JavaScript arrays are a more flexible data structure that can be composed of heterogeneous objects.

The benchmark results are depicted in Figure 6.9. In this setting, we measure the median of 20 executions. As the figures show, DynQ outperforms implementations based on Lodash for all queries. In particular, DynQ outperforms Lodash with speedup factors ranging from 1.92x (MQ4) to 7.84x (MQ6). The high speedup on MQ6 is motivated by the fact that, similarly to the `data.table` API in R, also Lodash does not chain the filter with the limit operation, unlike DynQ. Moreover, DynQ performance are comparable with the hand written implementations in most of the queries. In particular, DynQ is slower than the hand written implementations only on MQ2 (0.91x), and faster on MQ6 (2.46x) and MQ7 (2.04x).

There are multiple reasons why DynQ is able to outperform the hand-written queries. First, the JavaScript semantics may enforce additional operations which are not required in data processing; as an example, JavaScript's Map performs hashing by converting each value into a string representation. Moreover, during the execution of hand-written queries, the JavaScript engine needs to perform more runtime checks than DynQ.

Besides performance, the implementations using DynQ are the most concise ones. In particular, the hand-written implementations of the micro-benchmark queries count 160 lines of code (LOC), the Lodash implementations count 58 LOC, and the DynQ implementations count 40 LOC.

Benchmarks on Existing Codebases. We evaluate DynQ on an existing code base by comparing the performance of an existing JavaScript library against an equivalent implementation that uses DynQ. In particular, we selected the npm module *cities* [17], which exposes a dataset of locations and offers an API for selecting and filtering elements. In this setting, we measure the median of 1000 executions (after a warmup of 5000 executions).

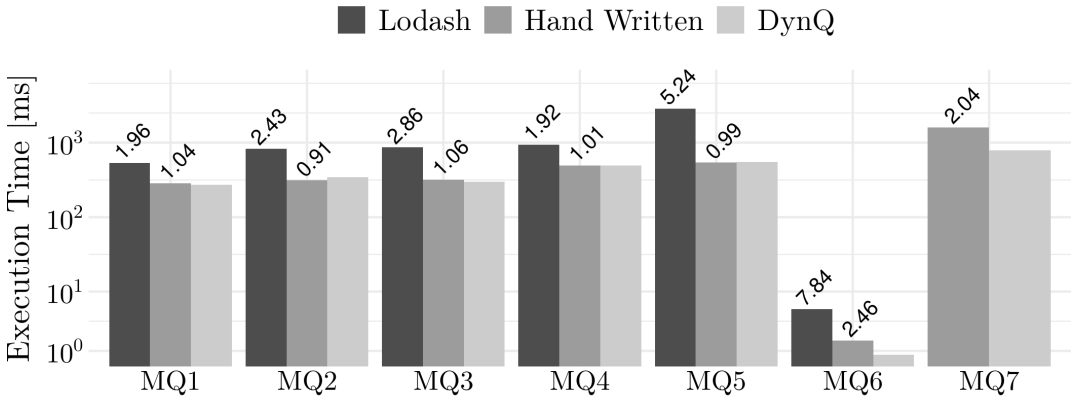

Figure 6.9. JS micro-benchmark (SF-1).

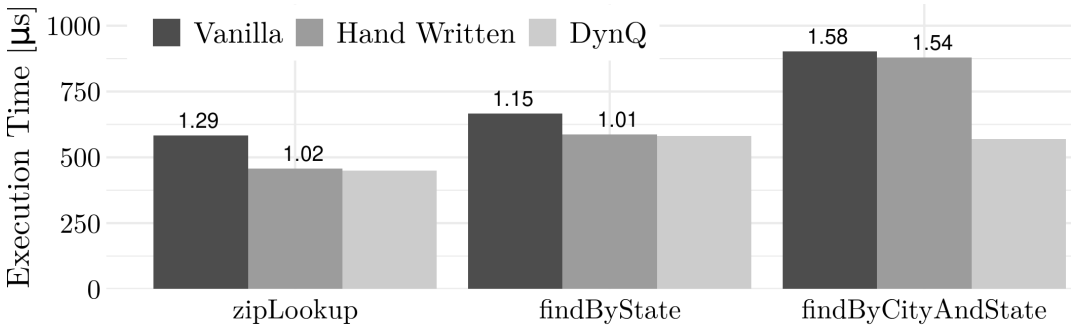

Figure 6.10. JS benchmark on cities module.

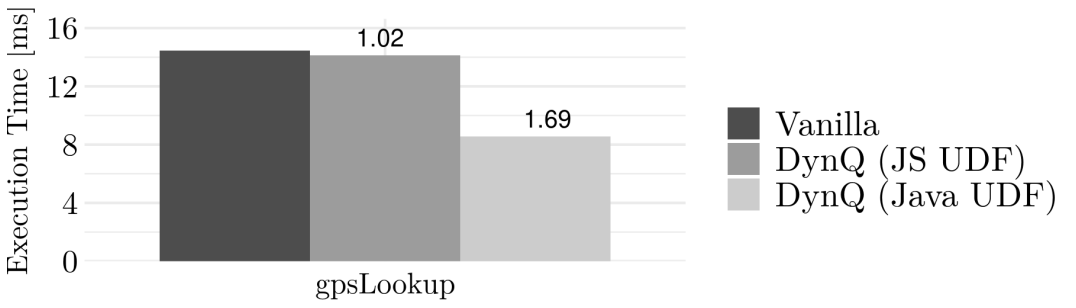

Figure 6.11. JS benchmark on cities module with UDF.

The npm module *cities* stores data in a single table (i.e., in a JavaScript array). The API offered by *cities* are listed below.

- `findByState`: finds the first location which matches a given state name.
- `findByCityAndState`: finds all the locations which match a given city name and state name.
- `zipLookup`: finds the first location which matches a given zip code.
- `gpsLookup`: finds the closest location of a given point (by latitude and longitude).

This module implements the first three API using Lodash, whilst the fourth API is manually implemented with hand-optimized code, which relies on the npm module *haversine* [41] for evaluating the distance between two points. Due to the simplicity of the API of the *cities* module, we also implemented an hand-optimized version of the first three API. We have not reimplemented the `gpsLookup` API, since the original version is already hand-optimized and it does not use any third-party data-processing library. For evaluating DynQ on the `gpsLookup` API, we use two versions; one version (DynQ(JS UDF)) uses the JavaScript module `haversine` as UDF for calculating the distance between two points, whilst the other version (DynQ(Java UDF)) uses a Java UDF instead of the JavaScript one. We manually implemented the Java UDF by carefully replicating the JavaScript version, such that the executed algorithm is exactly the same. As introduced in Section 5.3, we implemented all the API in cities with DynQ leveraging prepared statement.

The benchmark results are depicted in Figure 6.10 for the first three API, and in Figure 6.11 for the `gpsLookup` API. Since for the latter experiments we use DynQ in two different ways (i.e., implementing the UDF in JavaScript and Java), Figure 6.11 shows (above the bars of those two implementations) their respective speedups against the original implementation. As the figures show, DynQ outperforms both Lodash and the hand-optimized implementations in all API. Moreover, the evaluation of the `gpsLookup` API shows that evaluating an UDF with DynQ does not introduce any overhead when the UDF is implemented in the host dynamic language (i.e., JavaScript). This is expected, since, as discussed in Section 4.2, GraalVM can inline the machine code generated from the JavaScript UDF within the query execution code. Moreover, when the UDF is implemented in Java, performance improves, i.e., we measure a speedup factor of 1.69x. This is expected, since executing the JavaScript UDF requires more

type checks than executing the UDF in Java. Our evaluation on existing codebases shows that, besides data analytics, DynQ is also a promising library for server-side Node.JS applications that perform in-memory data processing.

6.3.3 Reusable Compiled Queries.

Here, we evaluate reusable compiled queries, our novel approach to query compilation discussed in Section 5.2. As a benchmark, we re-implemented two workloads, i.e., Scrabble and Mnemonics, which are part of the Renaissance [85], a well-known Java benchmark suite. The original Java implementation of these benchmarks evaluate the Java 8 Stream API, a data-processing API offered by the Java class library. In this setting, we implemented the benchmarks on JavaScript using Lodash and DynQ with different implementations of the fluent API: the per-query compilation approach (Section 4.5; DynQ-per-query in the figures), explicit parametricity (Section 5.1; DynQ-par in the figures), and reusable compiled queries (Section 5.2; DynQ-rcq in the figures).

As expected, implementations based on DynQ with explicit parametricity outperform all other implementations on both benchmarks, so we use this configuration of DynQ as a baseline. In particular, the bar plots show, above each bar, the speedup of DynQ with explicit parametricity against the implementation referred by that bar. However, as discussed in Section 5.2, explicit parametricity introduces limitations in term of modularity. On the other hand, reusable compiled queries do not suffer from this limitation, as they offer a classic fluent API. In particular, modifying an application to switch from a common data-processing library (e.g., Lodash) to DynQ with reusable compiled queries requires only minimal effort. The first goal of this evaluation is to show that both explicit parametricity and reusable compiled queries outperform Lodash and DynQ using per-query compilation approach. The second goal is showing that the enhanced modularity of reusable compiled queries introduces a very low overhead w.r.t. explicit parametricity.

Scrabble is a simulation of the well-known board game, which evaluates the list of words that results in higher score among a given list of 124 455 words. On Scrabble, the execution time is mostly dominated by the UDFs in charge of filtering words and evaluating their score. For this reason, in comparison to the slowest implementation, i.e., Lodash, DynQ with explicit parametricity obtains a moderate speedup of 1.25x. In comparison with DynQ with per-query compilation approach, explicit parametricity obtains a speedup of 1.23x, which is motivated by the fact that per-query compilation approach requires DynQ to start executing the query in interpreted mode even if the same query is executed mul-

tiple times, since the compiled code is not shared among multiple runs. Finally, in comparison with DynQ with reusable compiled queries, explicit parametricity obtains a minimal speedup of 1.04x, showing that the better modularity of reusable compiled queries does not impair performance.

Mnemonics uses Java streams to compute mnemonic phone codes [68]. Since Mnemonics uses only simple query operators, we implemented this benchmark also with the JavaScript simple (but performance-oriented) implementation of the query operators (JS-Push in the figure). Since Mnemonics is implemented as a recursive function which invokes two queries on each recursive step, we consider it a relevant application for evaluating reusable compiled queries. Once the recursion is close to its halting case, those queries are executed on very small arrays, meaning that high throughput of a single call is required to achieve high performance in the whole computation. As discussed in Chapter 5, DynQ approach to query execution without leveraging parametricity or reusable compiled queries, i.e., DynQ-per-query, is not suitable for those kinds of application, as creating a fresh AST for each query execution leads to executing the whole workload through interpretation most of the time.

The benchmark results are depicted in Figure 6.13. As expected, the slowest implementation is DynQ-per-query, which is outperformed by explicit parametricity by a factor of 6.92x. The speedup of DynQ with explicit parametricity and reusable compiled queries against Lodash (2.69x for DynQ-par) is because the Mnemonics benchmark uses the `flatMap` operator and Lodash implements that operator by materializing intermediate results. Our JavaScript implementations of the query operators are similar to the ones in Lodash, but the `flatMap` operator is implemented without materializing intermediate results, as in DynQ, which explains the performance improvement of our JavaScript implementations w.r.t. Lodash.

Although the JavaScript implementation of the query operators is conceptually very similar to those in DynQ, leveraging explicit parametricity leads DynQ to a speedup of 1.57x against the JavaScript implementation. Also reusable compiled queries outperform the JavaScript implementation, since, as discussed in Section 5.2, reusable compiled queries can guarantee that the sequence of operators composing a pipeline is fully de-virtualized. Finally, we note that explicit parametricity leads to a speedup of 1.15x in comparison with reusable compiled queries, a higher speedup w.r.t. the one observed on Scrabble. This is expected, since Mnemonics is a function which, for the benchmark input, executes 338 recursive calls for a single run of the benchmark, so the overhead of reusable compiled queries w.r.t. explicit parametricity is amplified in comparison to Scrabble, since each recursive call involves a query execution.

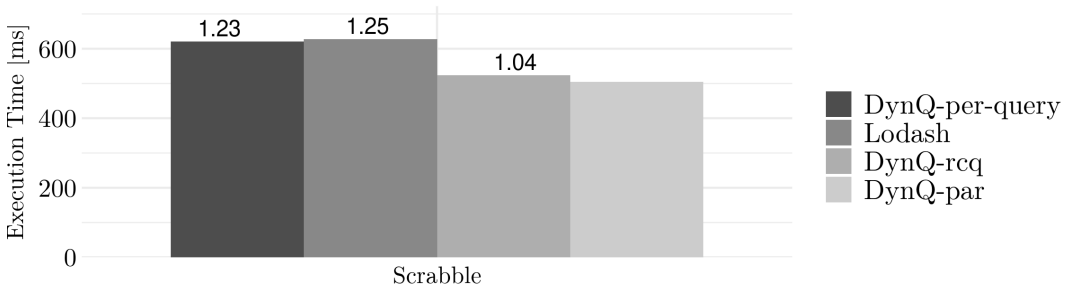

Figure 6.12. Scrabble benchmark on JavaScript with reusable compiled queries.

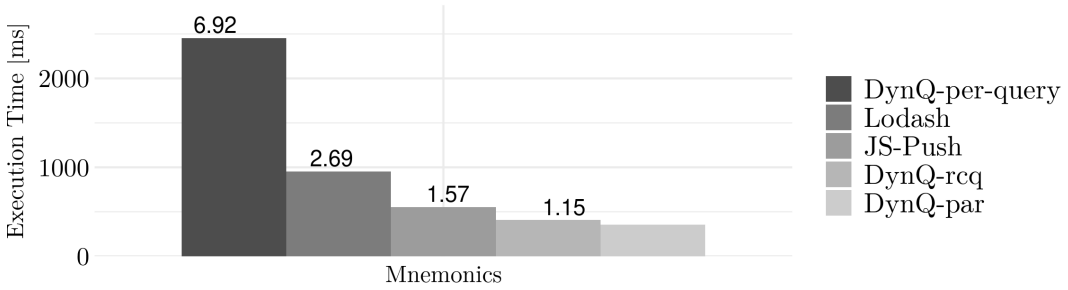

Figure 6.13. Mnemonics benchmark on JavaScript with reusable compiled queries.

Our evaluation on shared compiled queries shows that, besides high performance query execution, DynQ is now able to speed up high-throughput workloads, making DynQ an appealing drop-in replacement for data-processing libraries which offer a fluent API independently from the workload characteristics.

Chapter 7

Conclusion

Dynamically typed languages are nowadays the preferred solution adopted by developers and data-scientists for implementing analytical workloads, and they are also widely used for implementing server-side data-intensive applications. LINQ frameworks have been proposed as a declarative way to express queries on in-memory object collections as well as other data sources through LINQ providers. While LINQ support is currently missing for dynamically typed languages, we believe that developers and data scientists would benefit from declarative queries as well as other LINQ features. This dissertation bridged the gap between LINQ frameworks and dynamically typed languages.

7.1 Summary of Contributions

In this dissertation we introduced DynQ, a new query engine for dynamically typed languages. DynQ is based on a novel approach to SQL compilation, namely compilation into self-specializing executable ASTs. Our approach to SQL compilation relies on the Truffle framework and on GraalVM to dynamically compile query operators during query execution. Truffle was designed as a programming-language implementation framework; however, in DynQ we have exploited it in an innovative and previously unexplored way, i.e., as a code-generation framework integrated in a query engine. Moreover, we addressed the problem of optimizing high-throughout workloads, i.e., applications that perform data processing on many small datasets by introducing reusable compiled queries.

DynQ has been evaluated with two programming languages, namely R and JavaScript, against existing data-processing libraries and hand-optimized queries. Our evaluation shows that the performance of query evaluation with DynQ is comparable with, and sometimes better than, hand-optimized implementations,

outperforming existing data-processing systems and embedded databases in most of the benchmarks. Moreover, thanks to reusable compiled queries, DynQ is also able to outperform data-processing libraries also on high-throughput workloads that perform data processing on many small datasets.

To the best of our knowledge, DynQ is the first system which integrates a query engine within a polyglot VM directly interacting with its JIT compiler, and allowing the execution of federated queries on object collections as well as on file data and external database systems for multiple dynamically typed languages.

7.2 Discussion and Future Work

While DynQ has been designed targeting dynamically typed languages, its approach to query compilation and the proposed optimizations are not dedicated only to dynamically typed languages, and we believe that DynQ can bring benefits to statically typed languages as well, in particular to those executing in a virtual machine, e.g., Java. Those languages often offer dynamic features like inclusion polymorphism, which can make it impossible to statically deduce the exact type of the target for some operations. Indeed, JIT compilation is used within virtual machines to optimize also statically typed languages with dynamic features. For this reason, we believe that the query execution approach implemented in DynQ is suitable in the presence of such dynamic features. Indeed, as a future work, we plan to extend DynQ so that it can be used as a drop-in replacement for the Java Stream API.

We also believe that DynQ can also benefit data-processing scenarios where no dynamic feature is present. First, DynQ's approach to query processing based on hybrid interpreted-compiled execution to optimize workloads on both small and large datasets can be used also with a known schema. In this case, speculative assumptions and their guards used to type-check the input would not be needed. Moreover, we believe that such a dynamic approach to query execution can enable adaptivity on query engines based on query compilation, a feature which is commonly considered suitable only for engines based on query interpretation [51]. Adaptivity can benefit scenarios where there is no statistical information on the processed data. With our dynamic compilation approach, adaptivity can be obtained by recompiling part of the query-processing code depending on properties of the processed data which are not known before query execution; e.g., common adaptive query-execution techniques include runtime reordering of conjunctive predicates [42] as well as reordering of join operators [4].

Besides the features that DynQ offers to end users, we believe that DynQ would also be a useful framework in other data-processing domains. Indeed, since DynQ is a language-agnostic data-processing framework, and because of its flexibility in executing queries on different data representations, DynQ could be exploited for implementing query execution in the context of other existing data-processing frameworks written in any language supported by GraalVM. As an example, DynQ could be used for implementing query execution on external files with possibly malformed data (e.g., JSON files) in a similar way as done in our previous work on Spark SQL [95], i.e., by leveraging speculative optimizations.

We also believe that DynQ could be used within the GraalVM platform for implementing data-processing operators offered by the language, e.g., the R data-frame API, or the list- and set-comprehensions in Python. In particular, we believe that DynQ could be generalized and integrated within the GraalVM/Truffle ecosystem as a building block for implementing language features related to data-processing operations. Indeed, current implementations of those operations may easily violate the principle of not repeating yourself, since many data-processing operations (e.g., map, filter, reduce) are available in almost any dynamically typed language, with very similar semantics. By creating a common set of building blocks for data-processing operations, language developers could reuse them, focusing on their compositions and language-specific details, rather then repeatedly implementing similar operations for different languages. At the same time, developers of query operators can focus on data-processing optimizations.

Besides improving modularity and code reuse, integrating DynQ as a building block for data-processing operations would immediately bring other benefits into the GraalVM/Truffle ecosystem, in particular it would enable federated query execution. As an example, the data frame of the Truffle implementation of R could be trivially extended to support a join operation with any other data source for which a DynQ provider is implemented, e.g., a JSON file.

Finally, besides the high impact that DynQ can provide being the first LINQ system for dynamically typed languages, we believe that DynQ is a great tool for further research. As an example, DynQ could be extended for implementing runtime predicate reordering in the area of adaptive data-processing, as mentioned above. Indeed, we think that releasing DynQ is an important contribution for the database and programming-language research communities. For this reason, we released DynQ as an open-source project, available at `https://github.com/usi-dag/DynQ-VLDB`.

Bibliography

[1] AfterBurner Team. AfterBurner TPC-H Benchmark, 2020. https://github.com/afterburnerdb/afterburner/blob/master/src/tpch/benchmark_tpch.js.

[2] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia. Spark SQL: Relational Data Processing in Spark. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 1383–1394, 2015.

[3] ASM.js Team. asm.js, 2020. https://http://asmjs.org.

[4] R. Avnur and J. M. Hellerstein. Eddies: Continuously Adaptive Query Processing. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 261–272, 2000.

[5] S. Babu, R. Motwani, K. Munagala, I. Nishizawa, and J. Widom. Adaptive Ordering of Pipelined Stream Filters. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 407–418, 2004.

[6] E. Begoli, J. Camacho-Rodríguez, J. Hyde, M. J. Mior, and D. Lemire. Apache Calcite: A Foundational Framework for Optimized Query Processing Over Heterogeneous Data Sources. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 221–230, 2018.

[7] V. Benzaken, G. Castagna, L. Daynès, J. Lopez, K. Nguyen, and R. Vernoux. Language-Integrated Queries: a BOLDR Approach. *WWW*, pages 1–16, 2018.

[8] G. Bierman, E. Meijer, and M. Torgersen. Lost in translation: Formalizing proposed extensions to C#. In *OOPSLA 2007*, pages 479–498, 2007.

[9] Bob Hayes. Programming Languages Most Used and Recommended by Data Scientists, 2019. https://businessoverbroadway.com/2019/01

/13/programming-languages-most-used-and-recommended-by-data-scientists/.

[10] P. A. Boncz, M. Zukowski, and N. Nes. MonetDB/X100: Hyper-Pipelining Query Execution. In *CIDR 2005*, pages 225–237, 2005.

[11] D. Bonetta and M. Brantner. FAD.js: fast JSON data access using JIT-based speculative optimizations. *Proceedings of the VLDB Endowment*, pages 1778–1789, 2017.

[12] D. D. Chamberlin, M. M. Astrahan, W. F. King, R. A. Lorie, J. W. Mehl, T. G. Price, M. Schkolnick, P. Griffiths Selinger, D. R. Slutz, B. W. Wade, and R. A. Yost. Support for Repetitive Transactions and Ad Hoc Queries in System R. *ACM Trans. Database Syst.*, pages 70–94, 1981.

[13] S. S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS Project: Integration of Heterogeneous Information Sources. In *IPSJ*, 1994.

[14] J. Chen, S. Jindel, R. Walzer, R. Sen, N. Jimsheleishvilli, and M. Andrews. The MemSQL Query Optimizer: A modern optimizer for real-time analytics in a distributed database. *Proceedings of the VLDB Endowment*, pages 1401–1412, 2016.

[15] J. Cheney, S. Lindley, and P. Wadler. A Practical Theory of Language-Integrated Query. *SIGPLAN Not.*, pages 403–416, 2013.

[16] A. Cheung, S. Madden, A. Solar-Lezama, O. Arden, and A. C. Myers. Using Program Analysis to Improve Database Applications. *IEEE Data Eng. Bull.*, 37(1):48–59, 2014.

[17] cities Team. cities - npm, 2020. https://www.npmjs.com/package/cities/.

[18] A. Crotty, A. Galakatos, K. Dursun, T. Kraska, C. Binnig, U. Cetintemel, and S. Zdonik. An architecture for compiling UDF-centric workflows. *Proceedings of the VLDB Endowment*, pages 1466–1477, 2015.

[19] A. Crotty, A. Galakatos, K. Dursun, T. Kraska, U. Çetintemel, and S. Zdonik. Tupleware: Redefining modern analytics. *arXiv preprint arXiv:1406.6667*, 2014.

[20] A. Crotty, A. Galakatos, K. Dursun, T. Kraska, U. Çetintemel, and S. B. Zdonik. Tupleware: "Big" Data, Big Analytics, Small Clusters. In *CIDR*, 2015.

[21] J. W. Davidson and S. Jinturkar. Aggressive loop unrolling in a retargetable, optimizing compiler. In *International Conference on Compiler Construction*, pages 59–73, 1996.

[22] C. Diaconu, C. Freedman, E. Ismert, P. Larson, P. Mittal, R. Stonecipher, N. Verma, and M. Zwilling. Hekaton: SQL Server's Memory-Optimized OLTP Engine. In *Proceedings of the International Conference on Management of Data*, SIGMOD, 2013.

[23] M. S. Divya and S. K. Goyal. ElasticSearch: An advanced and quick search technique to handle voluminous data. *Compusoft, An International Journal of Advanced Computer Technology*, pages 171–175, 2013.

[24] J. Duggan, A. J. Elmore, M. Stonebraker, M. Balazinska, B. Howe, J. Kepner, S. Madden, D. Maier, T. Mattson, and S. Zdonik. The BigDAWG Polystore System. *SIGMOD Rec.*, 44(2):11–16, 2015.

[25] C. Duta and T. Grust. Functional-Style SQL UDFs With a Capital'F'. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 1273–1287, 2020.

[26] C. Duta, D. Hirn, and T. Grust. Compiling PL/SQL away. *arXiv preprint arXiv:1909.03291*, 2019.

[27] ECMAScript Team. ECMAScript Language Specification - ECMA-262 Edition 5.1, 2020. `https://www.ecma-international.org/ecma-262/5.1/#sec-15.9.1.1`.

[28] K. El Gebaly and J. Lin. In-Browser Interactive SQL Analytics with Afterburner. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 1623–1626, 2017.

[29] H. Fang. Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. In *CYBER*, pages 820–824, 2015.

[30] C. Freedman, E. Ismert, P.-Å. Larson, et al. Compilation in the Microsoft SQL Server Hekaton Engine. *IEEE Data Eng. Bull.*, 37(1):22–30, 2014.

[31] J. Friesen. Processing json with jackson. In *Java XML and JSON*, pages 323–403. Springer, 2019.

[32] Y. Fu, K. W. Ong, Y. Papakonstantinou, and M. Petropoulos. The sql-based all-declarative forward web application development framework. In *CIDR*, pages 69–78, 2011.

[33] H. Funke and J. Teubner. Low-Latency Compilation of SQL Queries to Machine Code. *Proceedings of the VLDB Endowment*, 14(12):2691–2694, 2021.

[34] V. Gadepally, P. Chen, J. Duggan, A. J. Elmore, B. Haynes, J. Kepner, S. Madden, T. Mattson, and M. Stonebraker. The BigDAWG polystore system and architecture. In *2016 IEEE High Performance Extreme Computing Conference, HPEC 2016, Waltham, MA, USA, September 13-15, 2016*, pages 1–6. IEEE, 2016.

[35] G. Graefe and W. J. McKenna. The Volcano optimizer generator: extensibility and efficient search. In *Proceedings of IEEE 9th International Conference on Data Engineering*, pages 209–218, 1993.

[36] M. Grimmer, C. Seaton, R. Schatz, T. Würthinger, and H. Mössenböck. High-performance cross-language interoperability in a multi-language runtime. In *Proceedings of the Symposium on Dynamic Languages*, DLS 2015, page 78–90. Association for Computing Machinery, 2015.

[37] P. M. Grulich, S. Zeuch, and V. Markl. Babelfish: Efficient Execution of Polyglot Queries. *Proceedings of the VLDB Endowment*, 15(2):196–210, 2021.

[38] T. Grust, J. Rittinger, and T. Schreiber. Avalanche-Safe LINQ Compilation. *Proceedings of the VLDB Endowment*, 3:162–172, 2010.

[39] S. Gupta, S. Purandare, and K. Ramachandra. Aggify: Lifting the curse of cursor loops using custom aggregates. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 559–573, 2020.

[40] M. Hausenblas and J. Nadeau. Apache drill: interactive ad-hoc analysis at scale. *Big data*, pages 100–104, 2013.

[41] haversine Team. cities - haversine, 2020. `https://www.npmjs.com/package/haversine/`.

[42] J. M. Hellerstein and M. Stonebraker. Predicate Migration: Optimizing Queries with Expensive Predicates. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 267–276, 1993.

[43] U. Hölzle, C. Chambers, and D. Ungar. Optimizing Dynamically-Typed Object-Oriented Languages With Polymorphic Inline Caches. In *Proceedings of the European Conference on Object-Oriented Programming*, ECOOP, pages 21–38, 1991.

[44] U. Hölzle, C. Chambers, and D. Ungar. Debugging Optimized Code with Dynamic Deoptimization. In *Proceedings of the Conference on Programming Language Design and Implementation*, PLDI, pages 32–43. Association for Computing Machinery, 1992.

[45] S. Idreos, F. Groffen, N. Nes, S. Manegold, K. S. Mullender, and M. L. Kersten. MonetDB: Two Decades of Research in Column-oriented Database Architectures. *IEEE Data Engineering Bulletin*, 35(1):40–45, 2012.

[46] K. Jäger. JSINQ - A JavaScript implementation of LINQ to Objects. 2009.

[47] N. D. Jones. An Introduction to Partial Evaluation. *ACM Comput. Surv.*, pages 480–503, 1996.

[48] V. Josifovski, P. Schwarz, L. Haas, and E. Lin. Garlic: a new flavor of federated query processing for DB2. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 524–532, 2002.

[49] Json Server Team. json-server, 2020. `https://www.npmjs.com/package/json-server/`.

[50] M. Karpathiotakis, I. Alagiannis, T. Heinis, M. Branco, and A. Ailamaki. Just-In-Time Data Virtualization: Lightweight Data Management with ViDa. *CIDR*, 2015.

[51] T. Kersten, V. Leis, A. Kemper, T. Neumann, A. Pavlo, and P. Boncz. Everything You Always Wanted to Know About Compiled and Vectorized Queries but Were Afraid to Ask. *Proceedings of the VLDB Endowment*, pages 2209–2222, 2018.

[52] T. Kersten, V. Leis, and T. Neumann. Tidy Tuples and Flying Start: Fast Compilation and Fast Execution of Relational Queries in Umbra. *The VLDB Journal*, 2021.

[53] O. Kiselyov. The Design and Implementation of BER MetaOCaml. In *Functional and Logic Programming*, pages 86–102, 2014.

[54] O. Kiselyov, A. Biboudis, N. Palladinos, and Y. Smaragdakis. Stream fusion, to completeness. In *Proceedings of the 44th ACM SIGPLAN Symposium on Principles of Programming Languages*, POPL, pages 285–299, 2017.

[55] Y. Klonatos, C. Koch, T. Rompf, and H. Chafi. Building efficient query engines in a high-level language. *Proceedings of the VLDB Endowment*, pages 853–864, 2014.

[56] C. Koch, Y. Ahmad, O. Kennedy, M. Nikolic, A. Nötzli, D. Lupei, and A. Shaikhha. DBToaster: higher-order delta processing for dynamic, frequently fresh views. *The VLDB Journal*, 23(2):253–278, 2014.

[57] A. Kohn, V. Leis, and T. Neumann. Adaptive execution of compiled queries. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, pages 197–208, 2018.

[58] B. Kolev, C. Bondiombouy, P. Valduriez, R. Jiménez-Peris, R. Pau, and J. Pereira. The CloudMdsQL Multistore System. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 2113–2116. Association for Computing Machinery, 2016.

[59] T. M. Kowalski and R. Adamus. Optimisation of language-integrated queries by query unnesting. *Computer Languages, Systems & Structures*, 47:131–150, 2017.

[60] K. Krikellas, S. Viglas, and M. Cintra. Generating code for holistic query evaluation. In *ICDE 2010*, pages 613–624, 2010.

[61] S. K. Lam, A. Pitrou, and S. Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

[62] C. Lattner and V. Adve. LLVM: A compilation framework for lifelong program analysis & transformation. In *International Symposium on Code Generation and Optimization*, pages 75–86. IEEE, 2004.

[63] A. Y. Levy, I. S. Mumick, and Y. Sagiv. Query Optimization by Predicate Move-Around. In *VLDB*, pages 96–107, 1994.

[64] LINQ Team. Language Integrated Query (LINQ) provider for C# - Finance & Operations |Dynamics 365, 2020. https://docs.microsoft.com/en-us/dynamics365/fin-ops-core/dev-itpro/dev-tools/linq-provider-c.

[65] LINQ Team. LINQ to Objects (C#), 2020. https://docs.microsoft.com/en-us/dotnet/csharp/programming-guide/concepts/linq/linq-to-objects.

[66] Lodash Team. Lodash, 2020. https://lodash.com/.

[67] S. Mandl, O. Kozachuk, and J. Graupmann. Bring Your Language to Your Data with EXASOL. *Datenbanksysteme für Business, Technologie und Web (BTW 2017)*, 2017.

[68] Martin Odersky. State of Scala, 2011. http://days2011.scala-lang.org/sites/days2011/files/01.%20Martin%20Odersky.pdf.

[69] M. Masse. *REST API design rulebook: designing consistent RESTful web service interfaces*. " O'Reilly Media, Inc.", 2011.

[70] P. Menon, T. C. Mowry, and A. Pavlo. Relaxed Operator Fusion for In-Memory Databases: Making Compilation, Vectorization, and Prefetching Work Together At Last. *Proceedings of the VLDB Endowment*, pages 1–13, 2017.

[71] P. Menon, A. Ngom, L. Ma, T. C. Mowry, and A. Pavlo. Permutable Compiled Queries: Dynamically Adapting Compiled Queries without Recompiling. *Proceedings of the VLDB Endowment*, 14(2):101–113, 2020.

[72] MongoDB Team. The most popular database for modern apps |MongoDB, 2020. https://www.mongodb.com/.

[73] D. G. Murray, M. Isard, and Y. Yu. Steno: Automatic Optimization of Declarative Queries. In *Proceedings of the Conference on Programming Language Design and Implementation*, PLDI, pages 121–131. Association for Computing Machinery, 2011.

[74] F. Nagel, G. Bierman, and S. D. Viglas. Code Generation for Efficient Query Processing in Managed Runtimes. *Proceedings of the VLDB Endowment*, pages 1095–1106, 2014.

[75] T. Neumann. Efficiently Compiling Efficient Query Plans for Modern Hardware. *Proceedings of the VLDB Endowment*, pages 539–550, 2011.

[76] T. Neumann and M. J. Freitag. Umbra: A Disk-Based System with In-Memory Performance. In *CIDR*, 2020.

[77] NPM Team. npm |build amazing things, 2020. `https://www.npmjs.com/`.

[78] Oracle. Using Prepared Statements, 2021. `https://docs.oracle.com/javase/tutorial/jdbc/basics/prepared.html`.

[79] Oracle, Team Java. Stream (Java Platform SE 8), 2020. `https://docs.oracle.com/javase/8/docs/api/java/util/stream/Stream.html`.

[80] M. Owens. *The definitive guide to SQLite*. Apress, 2006.

[81] S. Palkar, J. J. Thomas, A. Shanbhag, D. Narayanan, H. Pirk, M. Schwarzkopf, S. P. Amarasinghe, M. A. Zaharia, and S. InfoLab. Weld : A Common Runtime for High Performance Data Analytics. In *CIDR*, 2017.

[82] M. Pantilimonov, R. Buchatskiy, R. Zhuykov, E. Sharygin, and D. Melnik. Machine Code Caching in PostgreSQL Query JIT-Compiler. In *2019 Ivannikov Memorial Workshop (IVMEM)*, pages 18–25, 2019.

[83] H. Pirk, O. Moll, M. Zaharia, and S. Madden. Voodoo - a vector algebra for portable database performance on modern hardware. *Proceedings of the VLDB Endowment*, pages 1707–1718, 2016.

[84] B. PostgreSQL. PostgreSQL, 1996. `www.PostgreSQL.org/about`.

[85] A. Prokopec, A. Rosà, D. Leopoldseder, G. Duboscq, P. Tůma, M. Studener, L. Bulej, Y. Zheng, A. Villazón, D. Simon, T. Würthinger, and W. Binder. Renaissance: Benchmarking Suite for Parallel Applications on the JVM. In *Proceedings of the Conference on Programming Language Design and Implementation*, PLDI, page 31–47. Association for Computing Machinery, 2019.

[86] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[87] M. Raasveldt and H. Mühleisen. Vectorized UDFs in Column-Stores. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management*, SSDBM, pages 1–12. Association for Computing Machinery, 2016.

[88] M. Raasveldt and H. Mühleisen. Don't Hold My Data Hostage: A Case for Client Protocol Redesign. *Proceedings of the VLDB Endowment*, pages 1022–1033, 2017.

[89] M. Raasveldt and H. Mühleisen. DuckDB: an Embeddable Analytical Database. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD, pages 1981–1984, 2019.

[90] K. Ramachandra, K. Park, K. V. Emani, A. Halverson, C. Galindo-Legaria, and C. Cunningham. Froid: Optimization of imperative programs in a relational database. *Proceedings of the VLDB Endowment*, 11(4):432–444, 2017.

[91] J. Rao, H. Pirahesh, C. Mohan, and G. Lohman. Compiled Query Execution Engine Using JVM. In *22nd International Conference on Data Engineering (ICDE'06)*, ICDE, page 23. IEEE Computer Society, 2006.

[92] A. Rheinländer, U. Leser, and G. Graefe. Optimization of complex dataflows with user-defined functions. *ACM Computing Surveys (CSUR)*, 50(3):1–39, 2017.

[93] T. Rompf and M. Odersky. Lightweight Modular Staging: A Pragmatic Approach to Runtime Code Generation and Compiled DSLs. In *GPCE*, pages 127–136, 2010.

[94] V. Rosenfeld, R. Mueller, P. Tözün, and F. Özcan. Processing Java UDFs in a C++ environment. In *Proceedings of the 2017 Symposium on Cloud Computing*, SoCC, pages 419–431. Association for Computing Machinery, 2017.

[95] F. Schiavio, D. Bonetta, and W. Binder. Dynamic Speculative Optimizations for SQL Compilation in Apache Spark. *Proceedings of the VLDB Endowment*, pages 754–767, 2020.

[96] F. Schiavio, D. Bonetta, and W. Binder. Language-Agnostic Integrated Queries in a Managed Polyglot Runtime. *Proceedings of the VLDB Endowment*, 14(8):1414–1426, 2021.

[97] A. Shaikhha, M. Dashti, and C. Koch. Push vs. Pull-Based Loop Fusion in Query Engines. *Journal of Functional Programming*, 28, 2018.

[98] J. M. Smith, P. A. Bernstein, U. Dayal, N. Goodman, T. A. Landers, K. W. T. Lin, and E. Wong. Multibase: integrating heterogeneous distributed database systems. In *American Federation of Information Processing Societies: 1981 National Computer Conference, 4-7 May 1981, Chicago, Illinois, USA*, volume 50 of *AFIPS Conference Proceedings*, pages 487–499. AFIPS Press, 1981.

[99] Spectrum Team. Spectrum - Top Programming Languages, 2022. `https://spectrum.ieee.org/top-programming-languages-2022`.

[100] L. Spiegelberg, R. Yesantharao, M. Schwarzkopf, and T. Kraska. Tuplex: Data Science in Python at Native Code Speed. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 1718–1731, 2021.

[101] L. F. Spiegelberg and T. Kraska. Tuplex: robust, efficient analytics when Python rules. *Proceedings of the VLDB Endowment*, pages 1958–1961, 2019.

[102] Squel.js Team. Squel.js, 2020. `https://hiddentao.github.io/squel/`.

[103] StackOverflow Team. Stack Overflow Developer Survey 2019, 2020. `https://insights.stackoverflow.com/survey/2019/`.

[104] A. K. Sujeeth, K. J. Brown, H. Lee, T. Rompf, H. Chafi, M. Odersky, and K. Olukotun. Delite: A Compiler Architecture for Performance-Oriented Embedded Domain-Specific Languages. *ACM Trans. Embed. Comput. Syst.*, 2014.

[105] A. K. Sujeeth, T. Rompf, K. J. Brown, H. Lee, H. Chafi, V. Popic, M. Wu, A. Prokopec, V. Jovanovic, M. Odersky, and K. Olukotun. Composition and Reuse with Compiled Domain-Specific Languages. In *ECOOP 2013*, pages 52–78, 2013.

[106] R. Y. Tahboub, G. M. Essertel, and T. Rompf. How to Architect a Query Compiler, Revisited. In *Proceedings of the International Conference on Management of Data*, SIGMOD, pages 307–322, 2018.

[107] R. Tan, R. Chirkova, V. Gadepally, and T. G. Mattson. Enabling query processing across heterogeneous data models: A survey. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 3211–3220. IEEE, 2017.

[108] TensorFlow Team. TensorFlow, 2020. `https://www.tensorflow.org/`.

[109] K. M. M. Thein. Apache kafka: Next generation distributed messaging system. *International Journal of Scientific Engineering and Technology Research*, pages 9478–9483, 2014.

[110] A. Torres, R. Galante, M. S. Pimenta, and A. J. B. Martins. Twenty years of object-relational mapping: A survey on patterns, solutions, and their implications on application design. *Information and Software Technology*, 82:1–18, 2017.

[111] TPC. TPC-H - Homepage, 2019. `http://www.tpc.org/tpch/`.

[112] V8 Team. V8 Engine, 2020. `https://v8.dev/`.

[113] M. Vogt, N. Hansen, J. Schönholz, D. Lengweiler, I. Geissmann, S. Philipp, A. Stiemer, and H. Schuldt. Polypheny-DB: Towards Bridging the Gap Between Polystores and HTAP Systems. In V. Gadepally, T. Mattson, M. Stonebraker, T. Kraska, F. Wang, G. Luo, J. Kong, and A. Dubovitskaya, editors, *Heterogeneous Data Management, Polystores, and Analytics for Healthcare*, pages 25–36, Cham, 2021. Springer International Publishing.

[114] S. Wanderman-Milne and N. Li. Runtime Code Generation in Cloudera Impala. *IEEE Data Eng. Bull.*, pages 31–37, 2014.

[115] E. Wcisło, P. Habela, and K. Subieta. A java-integrated object oriented query language. In *International Conference on Informatics Engineering and Information Science*, pages 589–603, 2011.

[116] C. Wimmer and T. Würthinger. Truffle: A Self-optimizing Runtime System. In *SPLASH*, pages 13–14, 2012.

[117] T. Würthinger, C. Wimmer, C. Humer, A. Wöß, L. Stadler, C. Seaton, G. Duboscq, D. Simon, and M. Grimmer. Practical Partial Evaluation for High-Performance Dynamic Language Runtimes. In *Proceedings of the Conference on Programming Language Design and Implementation*, PLDI, pages 662–676. Association for Computing Machinery, 2017.

[118] T. Würthinger, C. Wimmer, A. Wöß, L. Stadler, G. Duboscq, C. Humer, G. Richards, D. Simon, and M. Wolczko. One VM to Rule Them All. In *Onward!*, pages 187–204, 2013.

[119] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: Cluster Computing with Working Sets. In *HotCloud*, pages 1–10, 2010.