Master's Thesis Nr. 327

Systems Group, Department of Computer Science, ETH Zurich

A Test Suite for Rumble

by

Stevan Mihajlovic

Supervised by

Dr. Ghislain Fourny, Prof. Dr. Gustavo Alonso

October 1, 2020 - April 1, 2021

D INFK

Abstract

The increasing amount of data available to process in today's world led to emerging of engines for large-scale data processing such as Spark. In order to make querying more efficient, Rumble is an engine that automatically runs queries on top of Spark, using the JSONiq as declarative querying language instead of Spark API.

Rumble is still in beta version and requires a testing mechanism that would verify behavior of its implementation. JSONiq inherits 95% of its features from XQuery, its XML counterpart. QT3TS is Test Suite with over 30000 test cases designed for testing XQuery.

This work aims to create a Test Suite based on QT3TS and implement a Test Driver that is capable of executing it on top of Rumble in order to verify its implementation.

The work was carried out gradually through several phases of implementation. In each phase, we were improving the implementation of Test Driver and improving Rumble. We also have implemented the necessary XQuery to JSONiq conversion that was required to execute tests.

As the final outcome of the work, we have managed to produce JSONiq Test Suite that can be used to verify any implementation, not just Rumble. We have also implemented Test Driver for Rumble. We have made a significant impact on improving the implementation of Rumble engine.

Maybe group first three

I am not writing this so that you can pity me. I am writing to share the story of how I learned to be grateful.

If you are thinking to quit, do not do it! If you believe your life is miserable, it is not! Stop for a moment and look back at your life. Always look back at what you already have. Look at all the privileges you are entitled to but you are taking them for granted. You came to this world as a tabula rasa, and the world owes you nothing by childbirth. Whatever you already have in life, you should consider a blessing. If life taught me anything: In a blink of an eye, you could lose everything.

In the end, it does not matter what you will achieve. What matters is what kind of life you are going to live. Live every day like it is your last and you will see how simple life is. And if you wake up one day, hating your life and the avatar that you have become, gather strength to reset and start all over again. To everyone that did terrible things to me, I can only say thank you. Thank you for making me play a game called life in veteran mode. You made me stronger, you made me what I am today. I died and I was reborn more than 5 times, what is your excuse?

The list of people I am grateful to is endless, but I am limited to 2 pages. Whoever is left out should know that they still have a place in my heart:

- Nikola and Ivana - my brother and my mom, for being the only two people in the world for whom I would give my life

- Dusan Malinov - for implanting the idea of ETH in the first place

- Team 7 GameLab - for healing me after first-semester depression

- Daniel Keller - my boss at ABB who believed in my skills and capabilities that even when I did not believe in them

- Simon Weber - for pulling out more in me and not slowing me down

- Ex-wife - for teaching me what being manipulated actually means

- Ex-flatmate - for showing me how a toxic person can break a family

- Tommaso Ciussani - for saving me from abusive flatmate by sharing his room with me until I found a better place to live

- Costanza Importa - my teammate in CIL project for being human, having empathy and covering up for me when I lost my dad

- Regula Cinelli - for teaching me not to wait for time to pass in order to get something over with, but to picture yourself as a winner

- Svilen Stefanov - for being a true friend and my consigliere that I perhaps do not deserve

- Team Nignite - for heartwarming working environment and support

- Diana Ghinea - for seeing the value in me when I could not, for waking up desire and fire in me to be a better version of myself

- My dad Milutin - for telling the same stories 10 and more times over and over again, teaching me that life is about memories

Contents

Chapter 1

Introduction

The increasing amount of data available to process, as well as the ever-growing discrepancy between storage capacity, throughput and latency, has forced the database community to come up with new querying paradigms in the last two decades. Data became nested and heterogeneous (JSON), and is increasingly processed in parallel (Spark). In order to make querying more efficient and accessible, Rumble [MFI+20] is an engine that automatically runs queries on semi-structured and unstructured documents on top of Spark, using the JSONiq language.

JSONiq [jso20]is a functional and declarative language that addresses these problems with its most useful FLWOR expression, which is the more flexible counterpart of SQL's SELECT FROM WHERE. It inherits 95% of its features from XQuery, a W3C standard.

The XQuery/XPath 3.* Test Suite (QT3TS) [W3C13] provides a set of tests with over 30000 test cases designed to demonstrate the interoperability of W3C XML Query Language, version 3.0 and W3C XML Path Language implementations.

The high-level idea of this work is to implement a Test Driver that can directly use QT3TS in order to test and verify Rumble implementation.

Chapter 2

Background and Related work

In this chapter, we will introduce context on which our work is based. For full overview, we must familiarize the reader with the following concepts: Big Data, NoSQL, MapReduce, YARN, Spark, JSON, JSONiq and finally Rumble.

Test Driver itself will be built as a layer on top of Rumble. Because of the architecture which enables data independence, we do not need to know its under-laying structure. However, seeing the full architecture and having an overview will help us make decisions throughout this work.

2.1 Big Data

Big Data in today's world has a broad scope and several definitions. Here we will present a certain view of the Big Data on which Rumble was based. We can look at the data being "big" in following 3 dimensions [Fou18]:

- Volume - This term simply corresponds to the number of bytes that our data consists of. To have an idea of the scale, in Big Data, we are often looking at PB of data. Scientific centers such as CERN produce tens of PB of data annually. The information, the data in today's world brings value. Not only scientific centers but also big companies gather data, store it in their data centers and process it in order to extract this value.

- Variety - Data often comes in different shapes. The most familiar ones are Text - completely unstructured data, followed by data organized as Cubes, Tables, Graphs, or Trees on which we will mainly focus. Until 2000's, the world was mainly oriented towards Relational Databases for which they under-laying shape is Table. Main focus was on introducing normalization forms with the idea to avoid data redundancy. Then the tables would simply be joined using SQL as the query language via the foreign keys. However, starting from 2000's Relational Databases and SQL could not satisfy the needs of real-world scenarios. Often data is

from 01 Big Data - Introduction although I updated from 2020 lecture

3

unstructured, nested, values are missing etc. This trend led to NoSQL Databases. The main focus in NoSQL Database is to perform opposite and actually de-normalize the data. Looking at the table, we would now allow non-atomic values in a single cell or even missing values. Such a transition leads the data shape to transform from flat homogeneous tables to nested heterogeneous trees. Choosing the correct data shape is essential. What CSV and SQL were in a relational database, for tree-shaped data we have JSON and XML as a data format with JSONiq and XQuery as their respective querying languages.

- Velocity - Data, in the end, is physically stored on some medium drive. The 3 main factors of this under-laying medium drive are Capacity, Throughput and Latency. From mid 1950's until today, we have witnessed tremendous increase in all 3 factors. Capacity has increased by up to 200×10^9, throughput by 10×10^3 and latency by 8 times. This ever-growing discrepancy between factors has brought needs for parallelization and batch processing. Since a single medium drive has increased capacity much more than throughput, we need to read data from multiple medium drives simultaneously in parallel to obtain data fast enough. At the same time, to face the discrepancy between throughput and latency, we need to obtain data in batches. Thus, the need for systems that can perform parallel batch processing has increased.

maybe insert a picture from presentation???

In summary, traditional RDBMS, such as Oracle Database or Microsoft SQL Server, have focused on being compliant with ACID (Atomicity, Consistency, Isolation and Durability) properties. Such RDBMS with homogeneous tables are good when handling a small amount of data. However, if we need to scale the data massively, we need to turn to different technologies. These traditional RDBMS that use File Systems such as FAT32 or NTFS for physical storage are not sufficient.

from 02 Big Data - Lessons learn

NoSQL (Not Only SQL) Databases, on the other hand, are compliant with CAP (Capability, Availability, Partition tolerance) theorem. Examples of new NoSQL databases that have emerged are key-value stores (DynamoDB), column-oriented stores (HBase) and document stores (MongoDb). They often use Distributed File System as physical storage such as HDFS. Instead of traditional scaling up, by buying single high-performance hardware, the orientation is towards scaling out by buying a lot of cheap commodity hardware. Such scaling enables that hardware costs grow linearly with the amount of data. These concepts lead to building high-performance and scale-able frameworks such as Hadoop that can query and process distributed massive data in parallel.

from 03 Data in the Large - Object and Key-Value Storage4

2.2 Hadoop

Apache Hadoop [Whi15] is an open-source framework written in Java that is able to manage big data, store it and process it in parallel in a distributed way across a cluster of commodity hardware. It consists of 3 components:

- HDFS - Storage layer
- MapReduce - Processing layer
- YARN - Resource management layer

from 04 Data in the Large - Distributed File systems and Exercise03 HDFS

In this section, we will briefly introduce each of the layers. It will help the reader to better understand Spark in the upcoming chapter.

2.2.1 HDFS

Hadoop Distributed File System - HDFS [SKRC10] is a physical storage layer of Hadoop inspired by GFS [GGL03] written in Java. It is one of the most reliable FS for storing big data distributed on a cluster of commodity hardware.

In this section, we need to understand how HDFS physically stores big data on the machines. When we say big data, we are thinking at the scale of millions of PB files. This means that files are bigger than a single drive (medium). Therefore, in such a setting, the most suitable is block storage. Unlike a typical NTFS system with allocation units of 4 KB, the block size is by default 64 or 128 MB. It is chosen as a good trade-off between latency and replication. Transferring multiple blocks bigger than 4 KB will reduce latency and also reduce the network overhead. When it comes to replicas, each block has by default 3 copies in case of a failure.

The architecture is master-slave. Master is NameNode and it is in charge of storing the namespace. Namespace is a hierarchy of files and directories. Since the blocks are 64 or 128 MB, the metadata is also small. And since we are storing a rather small amount of very large files, the whole namespace can fit in the RAM of the NameNode. In addition to the namespace, NameNode knows the file to blocks mapping together with location of blocks and its replicas. The blocks are stored on DataNodes that act as slaves. When clients want to read/write the files, they communicate with NameNode only once to receive the locations meaning that NameNode is not the bottleneck.

Such an architecture allows potential infinite scalability just by adding DataNodes, meaning that hardware cost grows linearly with the increase of data. The single point of failure is NameNode, meaning that we have Capability and Partition tolerance at the cost of Availability from CAP theorem. In case of a failure, there is a secondary NameNode that would start-up. Also, it enables high durability with 3 replicas and read/write performance. When reading, it usually transfers a lot of data - batch processing.

2.2.2 MapReduce

MapReduce [DG04] in most broad definition is a programming model (style). It answers the question of how we process the data, and it consists of two crucial steps map and reduce alongside with shuffle as the intermediate step:

from 06 Data in the Large - Massive Parallel Processing

- Map - Input data is mapped into an intermediate set of key-value pairs

- Shuffle - All key-value pairs are shuffled in a way such that all pairs with the same key end up same machine

Maybe insert picture left to right Map Sort Partition Reduce

- Reduce - Data is aggregated on the machine and the output is produced

Example - Count occurrences of each word in a document of 1000 Pages:

1. What we can do is that we can first have a single map task per page. This way, those 1000 pages can be done in parallel. The map task will perform Map $(K1, V1) -> $ List $(K2, V2)$, where K1 is in the range from 1 to 1000 (for each page) and V1 is the text on each page. K2 will have values in the range of all possible words that occur in the document. V2 will always be 1. Such a mapper is very primitive. If reduce task is a function that is commutative and associative, then it is allowed to execute the same function in the map task to reduce the amount of shuffle that will happen afterward. As count is such a function, in the map task, we can already perform sum per key. It means that K2 stays the same and V2 will be the actual count per page!

2. As not all possible words will appear in all pages, we will simply put together collection of all the produced key value pairs and sort them per key. We will then assign all key-value pairs with the same key to the single reducer and partition the data accordingly.

3. Reduce task will perform $(K2,$ List $(V2)) -> $ List $(K2, V3)$ - Reducer can output the same key value pair, but in general it can be any other. Finally V3 will be the sum of occurrences of the word K2!

In general MapReduce as programming model can be used in any framework with any under-laying physical storage such as Local File System, S3, Azure, HDFS. Here we will describe infrastructure in Hadoop version 1 where it is running on top of HDFS where we also have Resource Management layer. It is again master-slave architecture where we have JobTracker and Task-Tracker. JobTracker is the master with responsibilities of Resource Management, Scheduling, Monitoring, Job lifecycle and Fault-tolerance. One Job contains from multiple tasks, depending on how the data is split. And 1 task can be map or reduce task. 1 or more tasks are then assigned to TaskTracker that need to execute them. JobTracker is collocated with NameNode and TaskTracker usually with the DataNode in order to bring query to the data.

2.2.3 YARN

Yet Another Resource Negotiator - YARN [VMD$^+$13] is a Resource Management layer in Hadoop Version 2. Comparing to Version 1, we might notice that JobTracker has a lot of responsibilities. It is responsible for both types of jobs, scheduling and monitoring ones. In such a setting, JobTracker is acting as the "Jack of all trades" and becoming a bottleneck. This introduces scaleability issues such Hadoop could not handle more than 4000 nodes executing more than 40000 tasks (remember that job comprises a set of task).

This is the reason of introducing YARN. YARN clearly separates scheduling and monitoring responsibilities. The architecture is again master-slave where we have ResourceManager and NodeManager. There is a single ResourceManager per cluster that is in charge of only scheduling and performs: Capacity guarantees, Fairness, SLA, Cluster Utilization, Assigns containers. It has global overview of all cluster resources and provides leases for containers. One node in a cluster has one NodeManager and many Containers. Container is abstraction in which task can be run and it comprises a set of resources such as RAM, CPU, Storage, Bandwidth that can be allocated to ApplicationMaster. ApplicationMaster has the responsibility to handle monitoring. In particular it is in charge of: Fault tolerance, Monitoring, Asking for resources, Tracing job progress/status, Heartbeating to resource manager, Ability to handle multiple jobs. We have many ApplicationMasters in a cluster, each job has 1 application master, but not every node has to have a ApplicationMaster. In essence it can happen that single node has multiple ApplicationMasters, each responsible for different job completely unaware of the existence of other ApplicationMasters on the node. Finally it is should be noted that ApplicationMaster is a container. Described architecture solves the bottleneck issue allowing cluster to scale up to 10000 nodes and 100000 tasks.

> from 07 Data in the Large - Resource Management and Exercise 06 Spark

Full flow of duties overview:

- Clients submits a job.

- ResourceManager creates a job and returns ID.

- Client sends its requirements.

- ResourceManager tells a NodeManager to promote one of containers to ApplicationMaster.

- ResourceManager tells maximum capacity of containers.

- ApplicationMaster requests containers.

- ResourceManager assigns containers

> Maybe architecture picture with containers and all

YARN offers couple of types of schedulers that based on application and its request in terms of resources perform allocation.

2.3 Spark

Apache Spark [ZXW+16] [CZ18] [KKWZ15] is is an open-source engine for large-scale data processing. We see it as generalization of Map Reduce. From straight pipeline of two tasks, map and reduce, it generalizes it to any Directed Acyclic Graph - DAG. DAGs are built around Resilient Distributed Datasets - RDDs [ZCD+12] which are abstraction for partitioned collection of values. On RDDs, we can perform creation, transformation and action. In Spark we need to make a clear separation of two plans, two graphs - lineage and DAG.

DAG is basically a physical plan of execution. A DAG is created when the user creates a RDD (by referencing a dataset in an external File System for example) and applies chains of lazy transformations on it. When action is called, it triggers the computation. The DAG is given to the DAG Scheduler which divides it into stages of tasks. A stage is comprised of tasks based on partitions of the input data. The Stages are passed on to YARN that now executes it physically. Since Spark has end to end DAG, it can figure out which tasks can be done in parallel. All these will then run into parallel on several nodes.

> from 08 Data in the Large - Massive Parallel Processing (SPARK) and Exercise 06 Spark

Lineage graph tells us a logical plan. It tells us which RDD originates from which RDD. All the dependencies between the RDDs will be logged in lineage graph, rather than the actual data. This is called lazy evaluation, it only gets triggered when action is called. This lineage is used to recompute the RDD in case of failure.

Fault tolerance using lineage - Imagine that we start with a RDD on which we need to perform couple of transformations and finally an action. Such RDD would first get partitioned so that it can be handled by multiple nodes. Imagine that some node fails, it means that only the partitions that were on that node have to be recomputed. And lineage graph is telling us exactly which set of transformation is needed to reconstruct the RDD.

> Maybe a picture of Lineage from P8 - Spark adding on side what is RDD1, RDD2 . . . RDD4 and how could they be distributed

DataFrame is high level abstraction of RDD's. It is logical data model that enables users to view and manipulate data independently of physical storage. DataFrames store data in collection of rows enabling user to look at RDD's as tables. They are nothing more than named columns like we had before. Therefore, we can use high level declarative language - Spark SQL to query the data without caring about under-laying physical storage.

The main problem with DataFrames is that heterogeneous data that we are encountering in tree data shapes cannot fit in DataFrame. All the de-normalization that enabled nested, missing values or values of different type will not work. Running Spark on such a DataSet results in Spark skipping and leaving to user to manually handle heterogeneous data. DataFrames are simply not the correct representation for the tree shaped data.

2.3.1 Apache Spark vs Apache Hadoop MapReduce

For emphasizing power of Spark, we have found a nice comparison with Hadoop MapReduce that can be separated in following categories:

- Performance - Hadoop MapReduce stores the output on the disk after each map or reduce task. Spark keeps everything in memory. Spark performs better if all data is stored in RAM. If RAM is full, Spark uses disk but overall it is better.

- Ease of use - Spark has compatible API for Python, Scala, Java. On the other hand, Hadoop MapReduce is written in Java and it is hard to learn the syntax for programming.

- Cost - Spark needs a lot of RAM so it is more expensive. All data needed for job has to fit in RAM

- Data processing - Spark can do graph, ML, batch and real time processing which makes it one platform for everything. Hadoop MapReduce is good for batch processing, but it doesn't support graph or real time processing.

- Fault tolerance - Hadoop MapReduce relies on hard drives. In case of failure, it can continue wherever it left of and save time. It also has replication for fault tolerance. Spark uses RDDs for fault tolerance. They can refer to any dataset in external storage like HDFS. If RDD is lost it is recomputed using transformations.

2.4 Querying Language

2.4.1 JSON

JavaScript Object Notation - JSON [JSO]is a text only, human-readable data format. It originates from JavaScript, but today it is a widely spread language-independent data format supported by many programming languages.

> Maybe a picture of a JSON document

As we have seen DataFrames in Spark and table data shape in general that can be stored in CSV data format, is not suitable for heterogeneous data and de-normalization does not work. On the other hand, tree data shape and JSON as data format in particular, is perfect choice for nested heterogeneous data. It supports nesting by using 2 structured data types:

- Object - collection of key-value pairs that acts as associative array (map) from string to any other type

- Arrays - ordered sequence of items of any type.

JSON also supports the 4 Atomic data types that can be String, Number, Boolean and Null.

2.4.2 JSONiq

JSONiq [FF13] as mentioned in the introduction is declarative and functional querying language created exactly to analyze files written in JSON data format. It is designed to analyze tree shaped data - nested and heterogeneous. It inherits 95% of its features from XQuery, its XML counterpart. It has data model that is able to capture all aspects of JSON data format.

We say it is declarative because user does not be aware of the under-laying structure. It is a query language like SQL is in the RDBMS, with a difference that it operates on JSON.

When it comes to data model, everything is expressed as a Sequence of Items. Item itself can be any of the 6 data types that JSON supports. In addition, Item can also be of a Function Type. Then all Expressions that exist operate only on Sequence of Items.

We say it is functional because Expression takes Sequence of Items as the input and as the output produces again Sequence of Items. This means that Expressions can be nested in any desired way.

The Expression can be:

> Check is this the correct List or create a better one such that it matches table for categorizing the runtime iterators

- Arithmetic
- Logic
- Comparison
- Literal
- JSON construction
- JSON navigation
- Sequence Construction
- Built-in function
- FLWOR expression.

FLWOR expression is the most powerful. Using its own clauses, it is capable everything Select From Where in SQL does - Selection, Projection, Grouping, Ordering, Join. In addition, that it can be nested any number of times in almost any order which SQL does not quite support. [Fou13]

> Maybe image with example queries

Tuple stream is produced by each clause in the FLWOR expression. It is a set of key-value pairs representing a binding from variable name to corresponding Sequence of Items. The clauses can consume these tuple streams and produce tuple streams. So between themselves, clauses communicate via tuple streams. As we said that all Expressions operate on Sequence of Items, only return clause that always has to be included in every FLWOR expression will actually consume tuple steam and produce Sequence of Items. [Fou13]

2.5 Rumble

Rumble is a query execution engine for large, heterogeneous, and nested collections of JSON objects built on top of Apache Spark [MFI$^+$20]. In this we will explain Rumble from user perspective, also mappings that were performed from JSONiq to Spark via Rumble and General Architecture of Rumble.

2.5.1 User Perspective

The user can use Rumble via command line or using the Rumble API for Java. The architecture overview is quite simple and presented in Figure 2.1. User only sees JSONiq query language and uses it to write desired query. Rumble then takes this query and it has logic capable to map and pass the query down to Spark. Spark is then able to execute query in the cluster. Spark usually reads from DFS, most typically HDFS we mentioned before. But more in general it can run on any FS or database. Typical input for a query is JSON Lines document. JSON Lines document uses JSON data format and the only difference from typical JSON document is that every line in the document is a single object. Such document has a bit lower human-readability for nested data compared to JSON document but it is quite commonly used in other fields such as Web Programming. [MF21]

| JSONiq Query | Rumble API / CL | RUMBLE | Mapping | Spark | HDFS |

Figure 2.1: Rumble Architecture Overview

2.5.2 Mapping

We said that Rumble has a logic that is capable to map query to Spark primitives. We also said that in JSONiq, everything is Sequence of Items. Therefore, Rumble uses interface Item in the code [IFM$^+$21]. All 6 types that were mentioned in Section 2.4.1 then implement this interface. After that, Item is wrapped using the Spark JavaRDD generic class and the mapping is complete! Spark is now able to execute queries using objects of the wrapper class.

Maybe I should present the full inheritance tree with subtypes as well

We also said that out of all Expressions, FLWOR Expressions are the most powerful ones and we can view them as set of clauses. Between themselves, clauses operate by consuming tuple streams instead of operating on Sequence of Items. Sequence of Items is produced only in the end with mandatory Return clause. Therefore, in the code [IFM$^+$21], Rumble uses class FlworTuple for wrapping to the Spark Dataset generic class that is used for DataFrames. For each clause, we have a RuntimeTupleIterator and they all, with exception of Return, have a reference to FlworTuple. More details in Subsection 2.5.3.

Or maybe I should just delete the whole Mapping since I have it at the end

2.5.3 General Architecture

So far, we were referring to Rumble as an engine. Essentially it is a compiler implemented in Java and as such it follows basic Compiler Design principles. In order not to break declarative property of JSONiq query language, it requires a proper separation of concerns. Irimescu in his thesis [Iri18] proposed the layered architecture described in Figure 2.2. It consists of 4 main phases:

1. Lexer and Parser take JSONiq query as an input and produce Abstract Syntax Tree - AST as the output

2. Translator takes the AST as the input and produces tree of expressions - Expression Tree as the output

3. Generator takes Expression Tree as input and converts it into tree of runtime iterators - Runtime Iterator Tree

4. Runtime iterators represent basically the code that can be executed on single node or on top of Spark

Figure 2.2: Rumble General Architecture

Lexer and Parser

The first steps in analyzing source code, in this case query written in JSONiq query language, are Lexical and Syntax analysis's performed by Lexer and Parser modules respectively. For rather simple languages, such as JSONiq is, these two modules can be automatically generated from grammar of language. Thus, Another Tool for Language Recognition - ANTLR v4 framework [PQ95] is used. ANTLR needs grammar (.g4) file with definitions of all language constructs as the input. For Rumble, JSONiq.g4 file was implemented and using it ANTLR auto-generated Parser and Lexer together with BaseVisitor (implements visitor pattern) Java classes. In the code, you can now use first Lexer class that takes JSONiq query stream as input and then pass it to Parser class which will generate AST and conclude the so called "front-end" part of compiler.

Translator

In general with compilers, AST cannot be used directly. As explained in [Cik20], JSONiq is functional language that is composed of expressions. Thus, higher-level abstractions - Expression Tree is needed. To achieve higher-level

abstractions, following classes had to be implemented. On top of inheritance tree, we have abstract class Node from which Expression and Clause classes are derived. Clause class is then used for for deriving all clauses of FLWOR Expression. For all other Expression types mentioned in Section 2.4.2, classes were derived from Expression class.

Second part of generating Expression Tree required specific implementation of BaseVisitor class generated by ANTLR. BaseVisitor is a generic class and its specific implementation - TranslationVisitor class wraps around Node class.

Third part of generating Expression Tree is Static Context class containing map between variable names and sequence types. Each expression has its own static context.

Using all these classes, it is then possible to generate Expression Tree as explained in [Iri18]:

"The visitor starts at the top level expression and then moves through all of the children passing along the current static context while doing three things:

1. For any expression that it visits, it sets the static context to be equal to the currently generated one.

2. For any variable reference, it checks that the variable name is present in the current static context, otherwise it throws an error (at compile time).

3. For any variable declaration it creates a new static context containing the new variable and sets the previously existing static context as parent."

Generator

So called "back-end" - last part of Compiler includes code generation where the intermediate code gets transformed into assembly instruction and finally machine instructions. For this step in Rumble, we are performing conversion from Expression Tree to tree of runtime iterators. As Rumble was written in Java, runtime iterators are in charge of executing operations which get converted to Java bytecode.

All RuntimeTupleIterator implement RuntimeTupleIteratorInterface while all other runtime iterators implement RuntimeIteratorInterface. Both interfaces are similar to java.util.Iterator interface with methods such as hasNext() and next(). Using next(), runtime iterators can iterate over Sequence of Items and return results one item at a time. In addition, next() method triggers computation of all children iterators by recursively calling next() method in them. Result of such implementation is "lazy evaluation" where results are compute only when demanded.

These two runtime interfaces operate on Dynamic Context containing mapping between variable names and actual sequences of items. Static Context is in charge of static type checking performed at compile-time while Dynamic Context is in charge of dynamic type checking performed at runtime.

As pointed out in [MFI+20] "The main novelty in the design of Rumble is the runtime iterators that can switch dynamically between different execution modes, and which encode the decision of which nesting level is distributed". In total there are 3 different execution modes. In local execution mode runtime iterators are executed on single node locally in Java and they do not push computation to Spark. The other two, RDD-based execution (which uses Spark's RDD interface) and DataFrame-based execution (which uses the DataFrame interface) are executed on top of Spark. They both push computation to spark when dataset is large and there is a clear advantage over local execution mode. The modes which runtime iterator supports is based on category of JSONiq Expression and it is presented in Table 2.1. The column "As in JSONiq" represents classification as we have seen it before in Section 2.4.2 where Built-in Function and Sequence Construction are omitted since they are spread across several categories. The columns L, R and D represent local, RDD-based and DataFrame-based execution mode respectively while + sign signifies that this mode is supported.

Category	Expression/Clause	As in JSONiq	L	R	D
local-only	(), {$k:$v}, [$seq], $$, +, -, mod, div, idiv, eq, ne, gt, lt, ge, le, and, or, not, $a\|\|$b, $f($x), $m to $n, try catch, instance of, castable, cast, some $x in $y satisfies...	Arithmetic, Comparison, Logic, Literal, JSON construction	+		
sequence-transforming	$seq[...], $a[$i], $a[], $a[[]], $o.$s, $seq!..., annotate, treat	JSON navigation	+	+	+
sequence-producing	json-file, parquet-file, libsvm-file, text-file, csv-file, avro-file, root- file, structured-json-file, parallelize			+	+
sequence-combining	seq1,$seq2, if ($c) then... else..., switch ($x) case... default..., typeswitch ($x) case... default...		+	+	+
FLWOR	for, let, where, group by, order by, count, return	FLWOR	+		+

Table 2.1: Runtime iterator categorization for JSONiq expressions and clauses

> Check if table makes sense for the plus signs because at the end of this section we have additional explanation which primitives are used (RDD union, flatmap etc...)

Local-only iterators executed in the local execution mode basically come down to implementing the Expression's behavior in Java. On the other hand, RDD and Dataframe-based execution modes require a mapping to Spark primitives as explained in Section 2.5.2. There is an essential difference between these two modes that are running on top of Spark. The Dataframe-based mode is used in case that internal structure is known statically. This mode is also preferred over RDD-based mode as it faster in execution. On the other hand, RDD-based mode is used whenever the structure is unknown.

Rumble in its initial version was using RDD-based mode for FLWOR Expressions. However, all FLWOR Clauses Iterators, with exception of Return, operate with FlworTuple. From the query it is possible to derive static type of the variables in the tuples and therefore represent them as columns in DataFrame. Today, the RuntimeTupleIterator is using SQL queries instead of RDD transformations of Spark. We will not explain in detail the new mappings for each and every FLWOR Clause but we will make a parallel to For Clause and reuse example from [MFI+20]. If the current variables are x, y, and z, and the new variable introduced by the for clause is i, then the for clause is mapped to the following:

SELECT x, y, z, EXPLODE(UDF(x, y, z)) AS i FROM input_stream

Spark's EXPLODE functionality corresponds to flatMap() on RDDs, while UDF is a Spark SQL user-defined function that takes all existing variables as input, and returns the resulting sequence of items as a List<Item>.

Chapter 3

XQuery/XPath 3.* Test Suite(QT3TS)

3.1 Analysis

In this chapter, we will discuss design decisions that we have made during the development of Test Driver. The core idea is to develop Test Driver completely independently from Rumble by maintaining the code outside of Rumble.

3.1.1 Programming Language

We view Rumble as black-box and the single point of communication with Rumble should only be via the Rumble Java public API. Therefore, we have decided to implement Test Driver as Java Console Application. Furthermore, Rumble is also written in Java. We decided to setup our Java Console Application project to have two modules - Test Driver and Rumble module. Rumble module is the branch in repository created for the purpose of this work [Mih20]. Making Test Driver module dependent on it, we are allowing possibility to directly use Rumble and its classes in case that not everything is possible to be achieved by treating Rumble as the black-box.

3.1.2 Data Format

The XQuery/XPath 3.* Test Suite (QT3TS) is publicly available at W3C Public CVS Repository under module name 2011/QT3-test-suite [W3C11]. Since April 1st 2019, CVS tree has been discontinued and the repository has been migrated to W3C Public GitHub repository [W3C20]. The tests are published as a set of files - test sets containing in total more than 30000 test cases written mostly in XML format. W3C does not supply a Test Driver for executing the tests. Instead, for each implementation a Test Driver should be written. As these test sets are mostly written in XML format, the first component that our Test Driver will require is the XML parser.

3.1.3 XML Parser

XML parser is a program that allows our application to read and write XML documents. For our work, we have investigated following possibilities:

- DOM (Document Object Model) - This parser loads entire XML document in memory and uses interfaces to access the information. It can access couple of item elements at same time. It can be used for both reading and writing.

- SAX (Simple API for XML parsing) - This parser doesn't load XML document in memory. Instead, it allows us to register a handler with SAX parser. When parser goes through file it keeps invoking methods on the handler class for each item. It process it in sequence one at a time. For each new item it reads, it forgets state of previous items. Therefore, on each read we need to take appropriate action in our application. It is read only and also known as push parser. There is no handler on XML document side, only in our application.

- STAX (Streaming API for XML parsing) - This parser allows us to both read and write multiple documents at same time. Unlike SAX that reads one item at a time, STAX can be explicitly asked to get a certain item from XML document without loading it in memory. Therefore, we can look at it as mixture of DOM and SAX. It is pull parser and has handler on XML document as well

- JAXP (JAVA API for XML parsing) - Since JDK 1.5, the JAXP API has been available as a standard part of the Java platform, and it provides access to XSLT transformation, schema validation, and XPath processing services.

- Saxon [Kay20] - Open Source XSLT & XQuery processor developed by Saxonica Limited. The Saxon package is a collection of tools for processing XML documents. The main components accessible via API are:

  1. XSLT processor. Saxon implements the XSLT 3.0 Recommendation. The product can also be used to run XSLT 2.0 stylesheets, or XSLT 1.0 stylesheets in backwards compatibility mode.

  2. XPath processor. This supports XPath 2.0 and XPath 3.1. It can also be used in backwards-compatibility mode to evaluate XPath 1.0 expressions.

  3. XQuery processor. This supports XQuery 3.1, which also allows XQuery 1.0 or 3.0 queries to be executed.

  4. XML Schema Processor. This supports both XSD 1.0 and XSD 1.1. It can be used to support the schema-aware functionality of the XSLT and XQuery processors.

For parsing XML, we have decided to use Saxon. One may argue that for all 4 listed components, Java also has its own API – JAXP for 1st, 2nd and 4th together with XQJ for 3rd. However, in practice, Saxon is easier to use and more flexible than JAXP. Apart from that, main arguments are:

1. Saxon itself is one of the implementations for which Test Driver was also implemented. Based on Results Report [Kay16], it passes more than 99,9% of the QT3TS tests and it is considered a reference for XML.

2. Saxons implementation of the Test Driver can be used as a baseline for developing our own Test Driver.

3.2 Phase 1 Implementation

3.2.1 Description

In the first phase of the implementation we have analyzed the structure of QT3TS. We had to understand the under-laying structure of each and every test case. We had to see under which tags the information is stored in order to obtain it using Saxon API. Example test case in XML format:

```xml
<test-case name="fn-absint1args-1">
   <description>
      Test: absint1args-1 The "abs" function
      with the arguments set as follows:
      $arg = xs:int(lower bound)
   </description>
   <created by="Carmelo Montanez" on="2004-12-13"/>
   <environment ref="empty"/>
   <test>fn:abs(xs:int("-2147483648"))</test>
   <result>
      <all-of>
         <assert-eq>2147483648</assert-eq>
         <assert-type>xs:integer</assert-type>
      </all-of>
   </result>
</test-case>
```

The two most important tags in each test case are:

- Test - this is the test that should be executed on Rumble. It can be XSLT, XPath or XQuery expression.

- Result - this is the expected result outcome of the test tag. As it can be seen in the provided example, there are several types of assertions that we need to verify.

Test Driver's Test Case Handling Logic is supposed to iterate over catalog.xml using the Saxon API. This XML document contains list of all test-sets. Again, using the Saxon API, we iterate over test-cases in each of the test-sets. For each test-case, we are asking explicitly Saxon XML parser to get items under Test and Result tags. To use Saxon API, we need to know the structure. But, once Test Case Handling Logic obtains information under Test tag, it passes it down "as is" to Rumble API in order to execute the query. Rumble API returns the result which is then passed down to Test Result Handling Logic.

Test Driver's Test Result Handling Logic is in charge of determining which assertion needs to be performed. Here we provide the list of possible assertions:

- assert-empty - This assertion requires result to return empty sequence

- assert - This assertion requires us to run another query in which obtained result will be used as parameter of the new query. For example:

```
<test>math:cos(math:pi() div 2)</test>
<result>
    <assert>abs($result) lt 1e-15</assert>
</result>
```

- assert-eq - It requires us to run another query in form of obtained result "eq" value under this tag

- assert-deep-eq - Similar to assert-eq but runs "deep-equal" query

- assert-true - It requires result to return single Boolean value True

- assert-false - Opposite of assert-true

- assert-string-value - It requires that each item in the obtained result sequence is type of String and also "eq" to the sequence under this tag

- all-of - It contains multiple different assert tags described in this list and it requires all of them to be fulfilled

- any-of - Similar to all-of but requires only one of them to be fulfilled

- assert-type - Requires to check if obtained result is instance of this tag

- assert-count - Requires obtained result sequence size to be equal to this

- not - It requires to execute nested assertion with a negation

- assert-permutation - Requires result sequence to be permutation of this

- assert-xml - Requires result to be a XML document matching this one

- serialization-matches - Requires serialization of result to match this

After the assertion is performed, we need to classify the results. The idea is to make statistics that are described in 3.2.3. With such a classification we would be able to improve Rumble by reporting bugs in its implementation.

3.2.2 Architecture

The overview of scenario described in 3.2.1 can be seen in Figure 3.1

Figure 3.1: Phase 1 Architecture Overview

3.2.3 Results

As explained in 3.2.1, result obtained via Rumble API was compared with the expected result by applying the correct assertion check. In case assertion passed, test-case was considered a Success and otherwise Fail. The block of code performing these operations was surrounded by try and catch. In case that test failed because the syntax was not completely JSONiq, it would throw a RumbleException or more generally an Exception - Crash. Important note is that due to time limits, only two assertions were not implemented. Thus, assert-xml and serialization-matches will always result in Crash. With this implementation, we were to be able to distinguish 3 possible scenarios:

1. Success - Test case succeed

2. Fail - Test case failed because of bug in Rumble

3. Crash - Test case failed because it is not compatible with Rumble

The report is generated as .csv file having test-sets as rows and total number of test-cases per scenario in the columns. In Table 3.1 we will present the aggregated sum over all rows in the .csv file:

Scenario	Total test-cases	% of all test-cases
Success	2330	7.8
Fail	2769	8.8
Crash	26421	83.4

Table 3.1: Phase 1 Results Overview

3.3   Phase 2 Implementation

3.3.1   Description

After generating Phase 1 Implementation report described in Table 3.1, we carefully examined our implementation and identified 4 major pain points:

- Unstable implementation of assertion which resulted in implementing proper way of result binding in Rumble

- Too many crashing tests which resulted in implementing converter

- Improving Test Driver implementation resulted in breaking previously implemented features. Therefore, regression tests were introduced

- Some tests were supposed to Crash with expected error codes and our granularity was not appropriate for distinguishing all test-cases

Result Binding

To better understand the issues we have encountered, we will provide following code snippet:

```
private boolean AssertEq(List<Item> resultAsList,
        XdmNode assertion) throws UnsupportedTypeException {
    String assertExpression = assertion.getStringValue();
    List<String> lines = resultAsList.stream()
        .map(x -> x.serialize()).collect(Collectors.toList());
    assertExpression += "=" + lines.get(0);
    List<Item> nestedResult = runQuery(assertExpression);
    return AssertTrue(nestedResult);
}

private boolean AssertStringValue(List<Item> resultAsList,
        XdmNode assertion) throws UnsupportedTypeException{
    String assertExpression = assertion.getStringValue();
    List<String> lines = resultAsList.stream()
        .map(x -> x.serialize()).collect(Collectors.toList());
    return assertExpression.equals(String.join(" ",lines));
}
```

If we examine the AssertEq implementation, we will notice that lines.get(0) assumes that obtained result is a single item and takes first one. It does not handle sequences! Furthermore, handling sequences was only possible for AssertStringEqual in case that our result is sequence of strings by performing string concatenation. All other assertions such as Assert, AssertEq, AssertDeepEq are not possible to be implemented. Finally, if we remember Assert example from Section 3.2.1, we will notice that we had to perform string replace of $result with actual result obtained from the Rumble API.

Thus, Rumble was extended to support result binding. The change was made in Rumble implementation itself. The only modification required in Test Driver was to instantiate a new RumbleConfiguration and also new Rumble instance for each test-case that requires result binding. Check code below:

```java
private boolean Assert(List<Item> resultAsList,
        XdmNode assertion) throws UnsupportedTypeException {
    String expectedResult = Convert(assertion.getStringValue());
    return runNestedQuery(resultAsList, expectedResult);
}

private boolean runNestedQuery(List<Item> resultAsList, String
    expectedResult){
    RumbleRuntimeConfiguration configuration =
        new RumbleRuntimeConfiguration();
    configuration.setExternalVariableValue(
    Name.createVariableInNoNamespace("result"), resultAsList);
    String assertExpression = "declare variable $result external;"
        + expectedResult;
    Rumble rumbleInstance = new Rumble(configuration);
    List<Item> nestedResult = runQuery(assertExpression,
        rumbleInstance);
    return AssertTrue(nestedResult);
}
```

The main concern of the new implementation was that performing many instantiates might cause the execution time to increase dramatically. However, after run-time increased only by 15seconds from 2minutes - only 12.5%.

Once the result binding was implemented, it allowed us to run the assert type also as a query instead of calling the publicly exposed methods of the Item class in the Rumble Java API. In Phase 1 implementation, we had a switch case for every possible type that Rumble Java API supports, making code difficult for future maintenance and extension with new supported types. With running assert type as "instance of" query , we managed to have a single point of conversion performed in the beginning and applied for both test case and the expected result. Within conversion we would discover the unsupported type errors without the need of second switch case to check whether Rumble's API Item class supports the type or not. Furthermore, the previously implemented switch case had unsupported type as default therefore hiding some types that were supported but not specified in the documentation. The mentioned conversion will be explained more in detail in Section 3.3.1.

The clean separation that was performed here initialized idea and was a base plan for XQuery to JSONiq conversion logic separation. In Section 3.4.1 we will describe Architecture that has separate application that takes XQuery as input, performs conversion and outputs JSONiq test suite. Such approach would make the Test Driver easily maintainable and extensible!

Converter

As seen in Table 3.1, we had less than 10% Success test-cases as almost all of them required conversion to JSONiq. Here we will document all the conversions that we have performed on both Test and Result tags in this Phase.

The first conversion that we have performed is between types. Both XQuery and JSONiq have simple(atomic) and complex(non-atomic) types.

The list of atomic types that is currently supported by Rumble was taken from official Rumble documentation [IFM+20b] and conversion was implemented accordingly. For all types that are not supported, our code throws UnsupportedTypeException.

Following 3 complex (non-atomic) types were handled by following conversion:

1. array(*) was replaced with array*

2. item() was replaced with item

3. map(string, atomic) was replaced with object

On the other hand, following 7 complex (non-atomic) types could not be converted and they all throw UnsupportedTypeException:

1. document

2. element

3. attribute

4. text

5. comment

6. processing-instruction

7. xs:QName

Other conversions that were performed:

1. true() was replaced with true

2. false() was replaced with false

3. INF was replaced with Infinity

4. array access via . was replaced with $$

5. ' was replaced with "

6. prefixes fn, xs, math, map, array were removed

Other items that were unsupported in Phase 2 were node(), empty-sequence() and xs:NOTATION together with all error codes that are not in Table 3.6 that was taken from [IFM+20a].

Regression Tests

During Phase 1, we were performing iterations with goal to overall improve Test Driver's implementation. The good metric while performing these iterations was total number of test-case Crashes. Our goal was to reduce those numbers as much as possible. This was mainly handled by making following changes: bug fixes, software enhancements, configuration changes. Creating this changes in software development can usually lead to creating new issues that were not present before or re-emergence of old issues. In these cases it is quite common that software development requires regression testing. Regression testing (rarely non-regression testing) is re-running functional and non-functional tests to ensure that previously developed and tested software still performs after a change. If not, that would be called a regression [Reg]. During iterations, it was noticed that our approach of fixing and improving the application is highly exposed to changes that require regression testing.

While performing iterations, we had to ensure that any further implementation would not break the test-cases that were passing before and at the same time not introduce new test-cases that are Crashing. Thus, for each iteration we have maintained log files of all Passed (Success + Managed) and Crashed test-cases. In every next iteration we have done two comparison between new and previous log files. We have performed a check that compared whether all the passed test-cases from the previous implementation were also contained in the new implementation or not and created "List of test cases that were passing before but not anymore". For Crashes, we did opposite check and created list of "Tests that were not crashing before, but are now and not in list above".

Handling Error Codes and better granularity

In the Section 3.2.1 we have described all the possible assertions for which we are able to verify whether the result of executed query matches the expected result. However, some of the test-cases are different. They are there to verify that certain query cannot be parsed or executed because it is not compliant with the XML language. To better understand we provide an example below:

```xml
<test-case name="Literals036">
   <description>
      Test for invalid decimal literal
   </description>
   <created by="Mike␣Rorke" on="2005-02-03"/>
   <test>65535032.001.01</test>
   <result>
      <error code="XPST0003"/>
   </result>
</test-case>
```

The issue with such query is that decimal number cannot have more than one decimal separation character (dot). Such a query should not parse and there is a correct error code that should be reported instead - XPST0003 - "It is a static error if an expression is not a valid instance of the grammar defined in."

The issue in Test Driver implementation is that when such a query is passed down to Rumble, it will cause a RumbleException which essentially breaks the execution of the code. Meaning that they need to be handled differently. Essentially, in case that executed query causes RumbleException, we need to check whether it has error codes as expected result. If not, it will be considered a Crash. If it has error code as expected result, we need to check whether it matches the RumbleException's error code. Often, test-cases have several error codes that they would accept under any-off assertion in the result tag. And this is exactly the check that we are performing. In case that we have a single match with any of the possible error codes, we consider test-case to be a Success. Otherwise, in case that any of the possible error codes matches the error codes that are in Table 3.6 we consider test-case to be a Failed. Otherwise it will go into category Unsupported Error. The two tags for which we are not verifying assertion but checking error codes are:

- error - the code is under attribute code and starts with X

- assert-serialization-error - similar as error but starts with S

The complete classification diagram is shown in Figure 3.2.
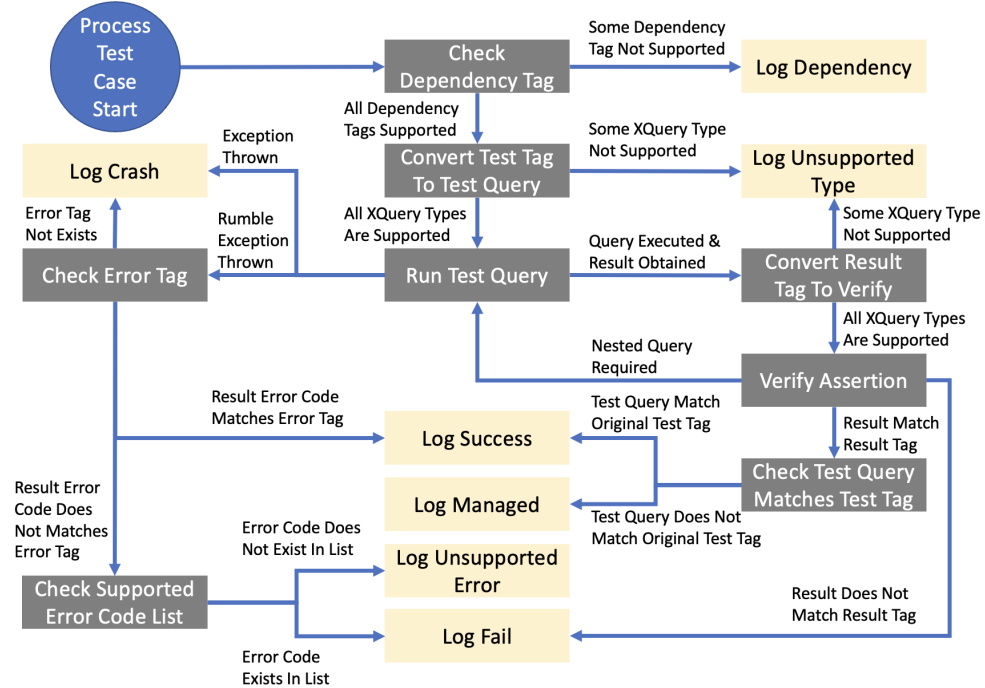
Figure 3.2: Phase 2 Classification Diagram

3.3.2 Architecture

The overview of scenario described in 3.3.1 can be seen in Figure 3.3
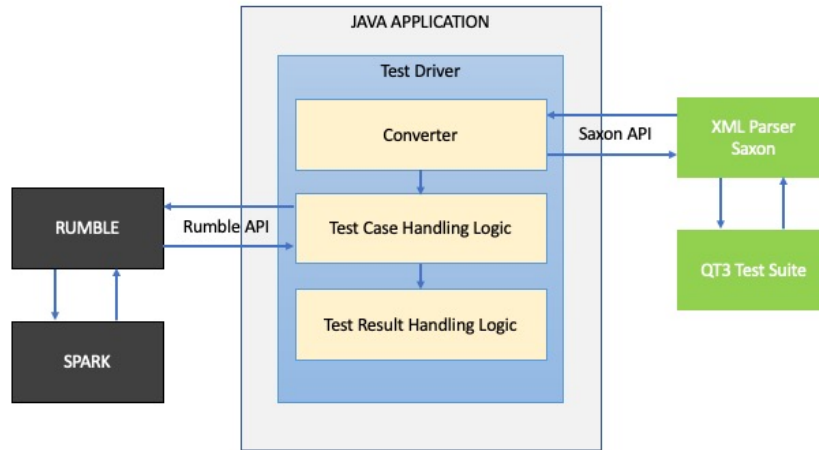
Figure 3.3: Phase 1 Architecture Overview

3.3.3 Results

As we have seen in Section 3.3.1, Crashes were not only capturing tests that are not JSONiq and needed conversion. They were also including the tests that could not succeed simply because Rumble does not support that type or error code. In Table 3.6 we can see the limited amount of supported types and error codes compared to XQuery W3C specification for which we are able to verify assertion. All others would then be ignored and classified differently. Furthermore, some of them were introducing dependencies. For example, in dependency tag it was possible to have request for particular version of XPath, XQuery or XSLT. While Rumble is backwards compatible with all versions of XPath and XQuery, it does not support XSLT. We have therefore created and divided test cases into 6 groups:

1. Success – Test that is passing the assertion and does not need Converter

2. Managed – Tests that would have failed assertion, but they were modified with hard-coded conversion into JSONiq using Converter

3. Failed – Tests that are failing because there is a bug in Rumble API or Test Driver implementation.

4. Dependency – Tests that are failing because dependency is not supported

5. Unsupported – Tests that are failing because type or error code is not supported yet

6. Crash – Any other exception

After introduction of the 6 above mentioned cases, together with small adjustments and bug changes, we were able to obtain:

Scenario	Total test-cases	% of all test-cases
Success	2686	8.52
Managed	4211	13.36
Fail	2554	8.10
Dependency	1481	4.70
Crash	13171	41.80
Unsupported	7412	23.52

Table 3.2: Phase 2 Results Overview

Managed category was introduced as it was identified that with simple hard-coded conversion we are able to obtain around 4200 passed tests increasing total percentage of passed tests by roughly 12%. At first, it seems that Success and Managed should be grouped into single category, but we decided to keep them separated. The reason behind is that while fixing bugs in both Rumble and Test Driver, we will increase the number of Success test cases. At the same time, we want to keep the track of Managed ones, because in Phase 3 Implementation we are planning to generalize the hard-coded conversion and create pure JSONiq Test Suite based on given XML ones.

For Skipped tests, these are the ones that it would not make sense to try to convert them to JSONiq. One example is XSLT tests and those should be skipped. We are keeping them in separate list as in Phase 3 Implementation we will skip from output and not include them in pure JSONiq Test Suite.

The main goal of performing iterations was to go through all the crashes and try to completely eliminate them. By doing so we would also improve the statistics by classifying them into other categories. At the same time, we were manually investigating test cases and trying to find the root cause. For some of them, our Test Driver implementation was improved. For some it was identified that the XQuery function was not yet supported by Rumle or it was having bugs so Rumble implementation was also improved. List of dependencies that were found in Test Suite were documented and classified according to Rumble documentation. The list is presented in Table 3.7 .

maybe some reference here, ask Fourny

Final goal was to identify test-cases that fail but can be converted to JSONiq. They could not be included into the automatic distinction of 7 above mentioned cases and had to be handled manually. They also helped with identifying what Phase 3 conversion also had to support.

3.4 Phase 3 implementation

3.4.1 Description

The main issue of Converter described in Section 3.3.1 was that it was hard-coded conversion using Java String.replace method. Such implementation can be very unstable. For example, we can look at 5th item of "other conversions" mentioned in Section 3.3.1 - replacing ' with ". For example, test-case Literals009 is verifying whether "test' is a valid String Literal. With our hard-coded conversion, we will make this test-case valid String Literal instead of it causing an Error Code XPST0003. Therefore, we have decided to implement Test Converter as separate module. It's main purpose is to generalize the hard-coded conversion. It would take QT3TS as Input and generate pure JSONiq Test Suite as output.

For implementing Test Converter we decided to create following classification of test-cases:

1. Fails, as expected and should not be converted to JSONiq. It will never be supported

2. Fails, as expected since it is not supported yet

3. Fails, but can be rescued with simple conversion. Any simple conversion like removing the "fn" prefix

4. Fails, but can be converted to JSONiq. Any complicated conversion like XML to JSON

5. Fails, because it is bug in Rumble

6. Succeeds

With this classification, we want to reuse most of Phase 2 Implementation Results presented in Table 3.2.

If we compare above described classification with classification in Table 3.2, we can notice that Item 5 corresponds to Fail, Item 6 to Success, Managed to Item 3. Item 2 to both Unsupported and Dependency.

Performing iterations in Phase 1, we want to distribute all Crashes into some of Item 4 or 1. Of course, it is in our interest to identify as many test-cases as possible as Item 4 and perform conversion in Test Converter. Everything that we cannot convert, we will classify as Item 1.

Items 1 will be excluded from Test Converter output. Items 2 on the other hand, will be excluded from Test Driver input. However, we also need to take into account that over time, as Rumble implementation improves, tests from Item 2 will be distributed into 4 other categories. Therefore, we want to make highly modular and extensible architecture. The important design decision remaining is the Data Format of the Test Converted output.

JSONiq and Test Converter Data Format

The JSONiq extension to XQuery allows processing XML and JSON natively and with a single language. This extension is based on the same data model as the core JSONiq and is based on the same logical concepts. Because of the complexity of the XQuery grammar, the JSONiq extension to XQuery has a less pleasant syntax that the JSONiq core. . When designing the Test Converter, we could have decided to use either XML or JSON as the underlaying language. However, as our Test Driver was already implemented in the previous phase and was expecting XML as input and using the before mentioned Saxon for parsing it, we have decided to keep the same language for output of the Test Converter.

`maybe cite something`

3.4.2   Architecture

The overview of scenario described in 3.4.1 can be seen in Figure 3.4
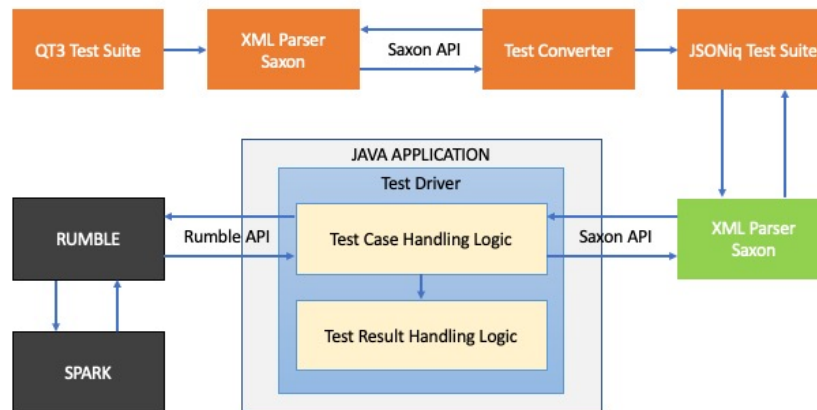
Figure 3.4: Phase 1 Architecture Overview

3.4.3   Results

After carefully analyzing the complete QT3TS, we have concluded that out of 424 test-sets QT3TS consists of, total of 143 can be classified of belonging in the Item 1 or Item 2 of classification documented in this phase of implementation. More specifically, we have assigned 61 test-sets to Item 1 presented in Table 3.4 and 84 test-sets to Item 2 presented in Table 3.5.

In addition to test-sets assigned to Item 2, we are adding the Unsupported and Dependency. Furthermore, several bugs were fixed in Rumble between Phase 2 and Phase 3 Implementation and results are presented in Table 3.3.

Scenario	Total test-cases	% of all test-cases
Item 1	3675	11.65
Item 2	12751	40.41
Item 3	7007	22.20
Item 4	4424	14.02
Item 5	999	3.17
Item 6	2701	8.56

Table 3.3: Phase 3 Results Overview

fn/base-uri.xml	prod/AxisStep.ancestor-or-self.xml
fn/doc.xml	prod/AxisStep.following.xml
fn/document-uri.xml	prod/AxisStep.following-sibling.xml
fn/element-with-id.xml	prod/AxisStep.preceding.xml
fn/generate-id.xml	prod/AxisStep.preceding-sibling.xml
fn/has-children.xml	prod/AxisStep.static-typing.xml
fn/id.xml	prod/AxisStep.unabbr.xml
fn/idref.xml	prod/BoundarySpaceDecl.xml
fn/innermost.xml	prod/CompAttrConstructor.xml
fn/in-scope-prefixes.xml	prod/CompDocConstructor.xml
fn/json-to-xml.xml	prod/CompCommentConstructor.xml
fn/lang.xml	prod/CompElemConstructor.xml
fn/name.xml	prod/CompNamespaceConstructor.xml
fn/namespace-uri.xml	prod/CompPIConstructor.xml
fn/namespace-uri-for-prefix.xml	prod/CompTextConstructor.xml
fn/nilled.xml	prod/ConstructionDecl.xml
fn/node-name.xml	prod/ConstructionDecl.schema.xml
fn/outermost.xml	prod/Comment.xml
fn/parse-xml.xml	prod/CopyNamespacesDecl.xml
fn/parse-xml-fragment.xml	prod/DirAttributeList.xml
fn/path.xml	prod/DirectConstructor.xml
fn/resolve-QName.xml	prod/DirElemConstructor.xml
fn/root.xml	prod/DirElemContent.xml
fn/xml-to-json.xml	prod/DirElemContent.namespace.xml
xs/token.xml	prod/DirElemContent.whitespace.xml
op/except.xml	prod/NameTest.xml
op/intersect.xml	prod/NodeTest.xml
op/is-same-node.xml	prod/SchemaImport.xml
prod/AxisStep.xml	prod/StepExpr.xml
prod/AxisStep.abbr.xml	prod/ValidateExpr.xml
prod/AxisStep.ancestor.xml	

Table 3.4: Item 1 - Fails, as expected and should not be converted to JSONiq. It will never be supported

fn/compare.xml	map/size.xml
fn/analyze-string.xml	map/put.xml
fn/collation-key.xml	map/remove.xml
fn/contains-token.xml	map/for-each.xml
fn/data.xml	array/append.xml
fn/default-collation.xml	array/filter.xml
fn/default-language.xml	array/fold-left.xml
fn/environment-variable.xml	array/fold-right.xml
fn/escape-html-uri.xml	array/for-each.xml
fn/filter.xml	array/for-each-pair.xml
fn/fold-left.xml	array/get.xml
fn/fold-right.xml	array/head.xml
fn/for-each.xml	array/insert-before.xml
fn/for-each-pair.xml	array/join.xml
fn/format-integer.xml	array/put.xml
fn/format-number.xml	array/remove.xml
fn/function-lookup.xml	array/reverse.xml
fn/function-arity.xml	array/sort.xml
fn/function-name.xml	array/subarray.xml
fn/implicit-timezone.xml	array/tail.xml
fn/iri-to-uri.xml	xs/dateTimeStamp.xml
fn/load-xquery-module.xml	xs/error.xml
fn/local-name.xml	xs/normalizedString.xml
fn/local-name-from-QName.xml	xs/numeric.xml
fn/namespace-uri-from-QName.xml	op/bang.xml
fn/parse-ietf-date.xml	op/QName-equal.xml
fn/parse-json.xml	prod/Annotation.xml
fn/prefix-from-QName.xml	prod/BaseURIDecl.xml
fn/QName.xml	prod/ContextItemDecl.xml
fn/random-number-generator.xml	prod/ContextItemExpr.xml
fn/resolve-uri.xml	prod/DefaultCollationDecl.xml
fn/sort.xml	prod/DefaultNamespaceDecl.xml
fn/static-base-uri.xml	prod/EQName.xml
fn/unparsed-text.xml	prod/ExtensionExpr.xml
fn/unparsed-text-available.xml	prod/ModuleImport.xml
fn/unparsed-text-lines.xml	prod/NamedFunctionRef.xml
map/merge.xml	prod/NamespaceDecl.xml
prod/MapConstructor.xml	prod/OptionDecl.xml
map/contains.xml	prod/OptionDecl.serialization.xml
map/find.xml	prod/UnaryLookup.xml
map/get.xml	prod/VersionDecl.xml
map/entry.xml	prod/WindowClause.xml

Table 3.5: Item 2 - Fails, as expected since it is not supported yet

Type	Status	Supported Error Codes
atomic	supported	FOAR0001
anyURI	supported	FOCA0002
base64Binary	supported	FODC0002
boolean	supported	FOFD1340
byte	not supported	FOFD1350
date	supported	JNDY0003
dateTime	supported	JNTY0004
dateTimeStamp	not supported	JNTY0024
dayTimeDuration	supported	JNTY0018
decimal	supported	RBDY0005
double	supported	RBML0001
duration	supported	RBML0002
float	not supported	RBML0003
gDay	not supported	RBML0004
gMonth	not supported	RBML0005
gYear	not supported	RBST0001
gYearMonth	not supported	RBST0002
hexBinary	supported	RBST0003
int	not supported	RBST0004
integer	supported	SENR0001
long	not supported	XPDY0002
negativeInteger	not supported	XPDY0050
nonPositiveInteger	not supported	XPDY0130
nonNegativeInteger	not supported	XPST0003
positiveInteger	not supported	XPST0008
short	not supported	XPST0017
string	supported	XPST0080
time	supported	XPST0081
unsignedByte	not supported	XPTY0004
unsignedInt	not supported	XQDY0054
unsignedLong	not supported	XQST0016
unsignedShort	not supported	XQST0031
yearMonthDuration	supported	XQST0033
		XQST0034
		XQST0038
		XQST0039
		XQST0047
		XQST0048
		XQST0049
		XQST0052
		XQST0059
		XQST0069
		XQST0088
		XQST0089
		XQST0094

Table 3.6: Rumble Supported Types and Error Codes

Dependency name	Status
higherOrderFunctions	supported
moduleImport	supported
arbitraryPrecisionDecimal	supported
schemaValidation	not supported (XML specific)
schemaImport	not supported (XML specific)
advanced-uca-fallback	not supported
non_empty_sequence_collection	not supported yet
collection-stability	not supported yet
directory-as-collection-uri	not supported yet
non_unicode_codepoint_collation	not supported
staticTyping	not supported yet
simple-uca-fallback	not supported
olson-timezone	not supported yet
fn-format-integer-CLDR	not supported yet
xpath-1.0-compatibility	not supported (XML specific)
fn-load-xquery-module	not supported yet
fn-transform-XSLT	not supported yet
namespace-axis	not supported (XML specific)
infoset-dtd	not supported (XML specific)
serialization	not supported yet
fn-transform-XSLT30	not supported yet
remote_http	not supported
typedData	not supported
schema-location-hint	not supported (XML specific)
calendar	not supported yet
unicode-version	supported
unicode-normalization-form	supported
format-integer-sequence	not supported yet
xsd-version	supported
xml-version	supported
default-language	only ën" supported
language	only ën" supported
spec	only "XT30+" not supported
limits	not supported yet

Table 3.7: Rumble Supported Dependency List

Chapter 4

Test Converter

In Figure 3.4 we have represented architecture that proposes implementing Test Converter as separate module. The main idea driving us was having a clean architecture that would make the Test Driver easily maintainable and extensible. Thus, we will separate these two and we can look at them as two different Java console applications. Test Driver will only handle executing the test-cases on Rumble while Test Coverter will only handle converting XQuery test suite to a JSONiq one.

However, In Section 3.4.1, we have explained the down sides of hard-coded conversion using Java String.replace method. Just having the clean architecture would not solve the hard-coding issues in Test Converter. Ultimately, we want to produce purely JSONiq Test Suite similar to QT3TS for XQuery. Such Test Suite would be available for everyone to use and verify other JSONiq implementations such as Zorba [Zor13], IBM WebSphere [Cor17], Xidel [Xid].

4.1 Rumble Architecture

First we need to think about achieving a clean XQuery to JSONiq conversion without hard-coding. Let us recall Rumble and its entire Section 2.5.3. We said that in order to finally execute the query, we are performing conversion from Expression Tree to tree of runtime iterators. The Expression Tree itself is a higher-level abstraction of JSONiq functional language that is composed of expressions. The entire Node Class Hierarchy that was implemented in Rumble is used to represent the Expression Tree.

If we are able to create Expression Tree from a query that is written in XQuery, it would mean that we have created JSONiq Expression Tree. Instead of creating the runtime iterator and executing the query, we can simply serialize Expression Tree to JSONiq query directly. We will reuse parts of Rumble and extends it's architecture so that it supports parsing queries written in XQuery.

This means that even though Test Converter and Test Driver will be two separate Java console applications, they will both use Rumble. One may argue that the success of standalone JSONiq Test Suite will in that case depend on Rumble implementation. And this is true, it will. But decision here is to use Rumble simply because using JSONiq Expression tree Rumble is the cleanest possible solution for conversion. Finally, it will also improve Rumble itself by supporting additional query language - XQuery.

In Figure 4.1 we have represented that will enable us to achieve the target. In the following sub sections we will explain the implementation details of XQuery Lexer & Parser and Translator together with Serialization part.
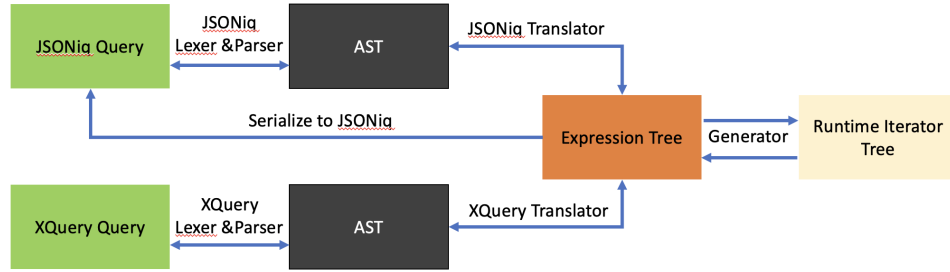

Figure 4.1: Rumble General Architecture with XQuery

4.2 Rumble Extension

Here we will go in details of implementation of each and every modification that has been performed in Rumble in order for us to be able to re-use the already existing classes in the Node Class Hierarchy.

4.2.1 Lexer and Parser

If we recall Section 2.5.3, for simple languages, these two modules can be automatically generated from grammar of language. XQuery is 95% similar to JSONiq, therefore we just need to create a grammar file similar to JSONiq.g4.

The initial approach was to implement the XQuery.g4 file ourselves based on already implemented JSONiq.g4 file in Rumble. However, XQuery is older than JSONiq and it was logical that such file was already implemented. Indeed, we managed to find an ANTLR4 implementation of xqDoc for XQuery [Xqu]. It is licensed under the same Apache License 2.0 like Rumble and we can re-use it. The grammar is stable and it was not changed for more than 10 months so we we do not expect it to be updated. Also it is compliant with XQuery 3.1 W3C Recommendation [W3C17].

Names of certain labels in grammar file were changed in order to match the code of TranslationVisitor class. which will be used as baseline to implement XQueryTranslationVisitor. The structure remains the same except for:

1. module was not according to the W3C Recommendation [W3C17] and allowed for multiple mainModules. See below the comparison between old and new module implementation.

   module : xqDocComment? versionDecl? xqDocComment? (libraryModule | (mainModule (SEMICOLON versionDecl? mainModule)*));

   module : xqDocComment? versionDecl? xqDocComment? (libraryModule | mainModule) ;

2. moduleDecl was changed to use uriLiteral instead of stringLiteral in order to match the implementation of TranslationVisitor class. This does not affect structure as uriLiteral comes down to stringLiteral in next level of nesting (uriLiteral: stringLiteral)

3. prolog was changed to use annotatedDecl in order to match the implementation of TranslationVisitor class. This does not affect structure as annotatedDecl offers the same options that were originally in prolog:

   annotatedDecl: varDecl | functionDecl | contextItemDecl | optionDecl;

4. functionDecl and varDecl were not according to the W3C Recommendation [W3C17]. It allowed for (annotations | ncName) while the ncName is under annotations but on 3 levels below in nesting

5. varDecl was not according to the W3C Recommendation [W3C17]. varValue and varDefaultValue were defined as expr instead of exprSingle and it allowed them to be surrounded by { }. See below the comparison between old and new varDecl implementation.

   varDecl: KW_DECLARE annotations KW_VARIABLE DOLLAR varName typeDeclaration? ((COLON_EQ varValue) | (KW_EXTERNAL (COLON_EQ varDefaultValue)?) | (LBRACE varValue RBRACE) | (KW_EXTERNAL(LBRACE varDefaultValue RBRACE)?)) ;

   varDecl: KW_DECLARE annotations KW_VARIABLE DOLLAR varName typeDeclaration? ((COLON_EQ varValue) | (KW_EXTERNAL (COLON_EQ varDefaultValue)?)) ;

   varValue: expr ; − > varValue: exprSingle ;

   varDefaultValue: expr ; − > varDefaultValue: exprSingle ;

6. squareArrayConstructor was changed to use expr instead of exprSingle (COMMA exprSingle)* in order to match the implementation of TranslationVisitor class. This does not affect structure as it is equivalent in next level of nesting (expr: exprSingle (COMMA exprSingle)* ;)

7. arrowExpr was changed to use complexArrow instead of arrowFunction-Specifier argumentList in order to match the implementation of TranslationVisitor class. This does not affect structure as it is equivalent in next level of nesting (complexArrow: arrowFunctionSpecifier argumentList;)

We have also implemented another script for handling the XQuery.g4, and using it, ANTLR auto-generated XQueryParser and XQueryLexer together with its corresponding XQueryParserBaseVisitor base class.

4.2.2 Translator

The second part requires us to implement XQueryTranslationVisitor that extends the generated XQueryParserBaseVisitor class and wraps around Node class. With proper implementation we would be able to achieve converting XQuery into the JSONiq Expression Tree. The implementation is mainly based on the already implemented JSONiq TranslationVisitor with modifications that we will document here.

Of course, not everything can be converted to JSONiq and it should not be converted. Below we will list the conversions that we have left out as they will never be supported:

- schemaImport
- copyNamespacesDecl
- constructionDecl
- boundarySpaceDecl
- optionDecl
- nodeComp
- unionExpr
- intersectExceptExpr
- parenthesizedExpr within the arrowFunctionSpecifier of arrowExpr
- URIQualifiedName
- validateExpr, extensionExpr within valueExpr of unaryExpr
- nodeConstructor within primaryExpr
- kindTest, typedMapTest, typedArrayTest within itemType
- axisStep within stepExpr
- multiple stepExpr within relativePathExpr
- single or double dash preceding relativePathExpr within pathExpr

In addition to that, there are conversions that we are yet not able to perform as the implementation is missing in Rumble. Unfortunately, here we are again introducing dependency on Rumble. But it was the best approach in order to perform the clean conversion without hard-coding. Following conversions are classified as out of scope of the thesis and will be supported in future once Rumble is upgraded:

> remove previous two sentences

- defaultNamespaceDecl

- decimalFormatDecl

- baseURIDecl

- contextItemDecl

- existUpdateExpr within exprSingle

- parenthesizedExpr and STAR object lookup within keySpecifier of lookup

- orderedExpr and unorderedExpr within primaryExpr

- windowClause within initialClause of flworExpr

- functionTest within itemType

For the versions that we support, since the grammar file is based on the XQuery 3.1 W3C Recommendation [W3C17], we have decided to support XQuery versions 1.0, 3.0 and 3.1 as version 3.1 is backward compatible.

For annotations, for now we only support single public annotations without prefix. Single since we can see from QT3TS in the prod/ModuleImport.xml, tests modules-pub-priv-29 to 36, we can see that "It is an error if a variable's annotations contains any combination of two annotations". In addition, according to the W3C Recommendation [W3C17] it is said that "If no prefix is present, the name is in the "http://www.w3.org/2012/xquery namespace". This results in following implementation of annotation handling:

1. Throw an error if there is more than 1 annotation in the annotations list

2. Otherwise, extract the EQName

3. Throw an error if prefix is not equal to http://www.w3.org/2012/xquery

4. Otherwise, throw an error if local name is not public

5. Otherwise, let the query with annotations to be converted

Namespaces fn and map that were removed in Section 3.3.1, we are now binding. Namespace xs is part of grammar. For namespaces map and array, no binding is needed as they are counted for Item 2 as mentioned in Section 4.3

Handling the Literals is slightly different. If we compare JSONiq with XQuery, we will see that true, false and null literals do not exist in XQuery. Instead they will be added as function calls.

The itemType of XQuery covers more possibilities compared to the JSONiq:

itemType: kindTest | (KW_ITEM LPAREN RPAREN) | functionTest | mapTest | arrayTest | atomicOrUnionType | parenthesizedItemTest ;

itemType : Kitem | atomicType; | Kobject | Karray | Kjson;

In JSONiq TranslationVisitor implementation, it was enough to call getItemTypeByName method and pass the parsed context. In XQuery implementation, this is only possible for the atomicOrUnionType. For all others, we have to handle it differently. First of all, kindTest will throw an error as it corresponds to the 7 complex (non-atomic) types that should not be converted (document, element etc.). The, functionTest is not yet supported in Rumble and it is out of the scope of thesis. The mapTest and arrayTest correspond to AtomicItemType.Object and AtomicItemType.Array item type of JSON if they are untyped, otherwise we throw an error. Item is the AtomicItemType.item while parenthesizedItemTest recursively calls the same method.

When it comes to arrayConstructor, there is a slight problem. JSONiq and XQuery do not have the same data model - there is a data model mismatch. XQuery allows us to have Sequence of Items as array members or object values. JSONiq does not support that and requires single items instead. The only way to store Sequence of Items in the JSONiq data model is using arrays. In order to overcome the dependency mismatch, we introduce mapping between two data models and we map nested Sequence of Items of XQuery to JSONiq array. In the case of squareArrayConstructor of XQuery, we recursively iterate over the exprSingle in expression and perform visitExprSingle. Each expression returned by visitExprSingle, is then wrapped into an array constructor. Finally all these array constructors are wrapped into Comma Expression that is then wrapped in the main array constructor. In the case of curlyArrayConstructor, we first obtain the content by calling visitExpr. Then we instantiate new Context Item Expression and wrap it into an array constructor. We combine these two by instantiating new Simple Map Expression that is then wrapped in the main array constructor.

4.2.3 Serialize to JSONiq

Instead of running the query and returning the iterator over the resulting Sequence of Items, we need to perform serialization. Serialization is basically inverted process of parsing. We have declared a new method serializeToJSONiq in the abstract class Node. All other classes that are used to create the JSONiq Expression Tree are derived from Node and in those classes we need to implement the method. We are now peeking into the XQuery.g4 grammar file in order to recreate the String representation of each and every Expression. In each implementation, we use the appropriate keywords and recursively call serializeToJSONiq for each nested Expression.

In Section 2.5.3, we have presented the 4 phases that Rumble goes through as a compiler. If we further analyze these phases through the code of Java Rumble API, we can identify following steps that are required in order to execute the query:

1. Instantiate Lexer from the input stream - this is the complete query

2. Instantiate Parser using the Lexer from Step 1

3. Instantiate Static Context (map between variable names and sequence types) using the URI file - this is the input file on which the query is executed

4. Instantiate Translation Visitor using Static Context from Step 3

5. Instantiate MainModule by calling visit method of Translation Visitor and passing the main module extracted by the Parser as the argument

6. In the runQuery method of Java Rumble API, first 5 Steps are performed by calling parseMainModuleFromQuery method that returns Main Module from Step 5

7. Instantiate Dynamic Context (mapping between variable names and actual sequences of items) using returned Main Module from Step 6

8. Instantiate RunTimeIterator using returned Main Module from Step 6

9. Return the Sequence of Items result instantiated using the Iterator from Step 8 and Dynamic Context from Step 7

To perform only serialization without obtaining results of the query, we have extended Java Rumble API with additional method serializeToJSONiq. This method performs the first same first 6 Steps described above. After that, we simply call the implementation of serializeToJSONiq of abstract Node class via Main Module from Step 6. We pass String Buffer as the argument which gets populated and then the Java Rumble API serializeToJSONiq method returns it as a String. Such implementation allowed Rumble to provide conversion out of the box via JSONiq Expression Tree.
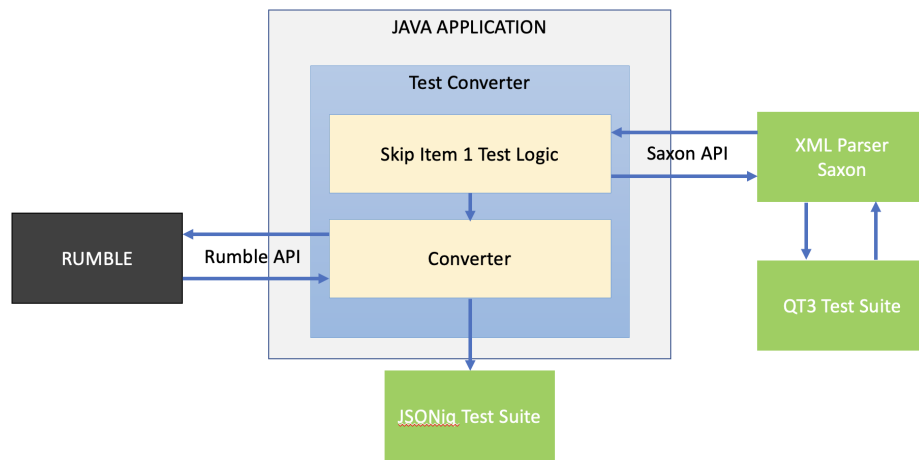
4.3   Architecture

As we said before in Section 3.4.1 we will distinguish 6 cases, but for this implementation we will need only first 2:

1. Fails, as expected and should not be converted to JSONiq. It will never be supported. - Documented in Table 3.4

2. Fails, as expected since it is not supported yet - Documented in Table 3.5

This separation is important to understand. As we want to produce standalone JSONiq Test Suite similar to QT3TS for XQuery, all Item 1 should not be included as they simply are not JSONiq! On the other hand, all Item 2 should be included because they can be used to verify other JSONiq implementations.

Final architecture of Test Converter can be seen in Figure 4.2. There is a small difference compared to Test Driver architecture. Test Converter does not need to execute the JSONiq query. Therefore as presented in Figure 4.1, we will simply call serializeToJSONiq instead of creating runtime iterators that would be executed on top of Spark i.e. we do not need Spark.

Figure 4.2: Test Converter Final Architecture

Final Test Driver architecture can be seen in Figure 4.3. We will maintain a list of Item 2 that will not be executed in Test Driver for Rumble. Test Driver will now take JSONiq Test Suite produced by Test Converter and execute it.
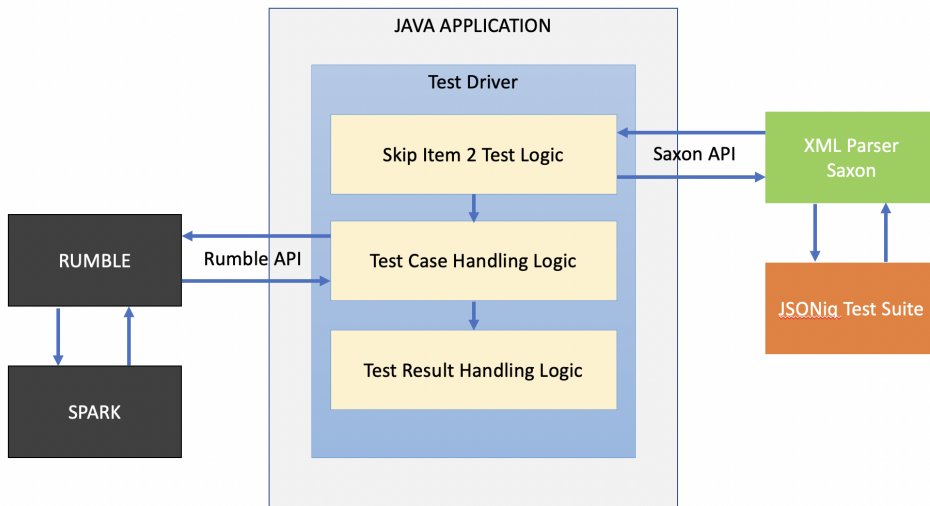
Figure 4.3: Test Converter Final Architecture

4.4 Implementation

The Converter part is using method binding provided by Saxon. First we define the ExtensionFunction and name under which it will be available in XQuery expression (convert). Secondly, we specify what is result of this function is (result of Java Convert method). Finally, we assign it a namespace (bf). We now define XQuery expression that has a function that recursively visits all nested XML tags. In case that they are test or result tags, it will pass the under-laying tag value to bf:convert and replace the tag value with returned result. We define external variable and pass the entire test-set to the XQuery expression. Check code below:

```java
Processor testDriverProcessor = new Processor(false);

ExtensionFunction converter = new ExtensionFunction() {
    @Override
    public QName getName() {
        return new QName(bindingNameSpace, "convert");
    }

    @Override
    public SequenceType[] getArgumentTypes() {
        return new SequenceType[]{
            SequenceType.makeSequenceType(
                ItemType.STRING, OccurrenceIndicator.ONE
        )};
    }

    @Override
    public XdmValue call(XdmValue[] arguments) throws SaxonApiException {
        String arg = arguments[0].itemAt(0).getStringValue();
        String result = Convert(arg);
        return new XdmAtomicValue(result);
    }

    @Override
    public SequenceType getResultType() {
        return SequenceType.makeSequenceType(
            ItemType.STRING, OccurrenceIndicator.ONE
        );
    }
};

testDriverProcessor.registerExtensionFunction(converter);
XQueryCompiler xqc = testDriverProcessor.newXQueryCompiler();
xqc.declareNamespace("bf", bindingNameSpace);
XQueryExecutable xqe = xqc.compile(
    "declare function local:transform($nodes as node()*) as node()*{\n" +
        "for $n in $nodes return\n" +
            "typeswitch ($n)\n" +
                "case element (test) " +
                    "return <test>{bf:convert($n/string())}</test>\n" +
                "case element (result) " +
                    "return <result>{bf:convert($n/string())}</result>\n" +
                "case element () " +
                    "return element { fn:node-name($n) } " +
                                "{$n/@*, local:transform($n/node())} \n" +
                "default " +
                    "return $n\n" +
    "};" +
    "declare variable $test-set external;\n" +
    "let $y := $test-set//test-set\n" +
    "return local:transform($y)");
XQueryEvaluator xQueryEvaluator = xqe.load();
xQueryEvaluator.setExternalVariable(new QName("test-set"), testSetDocNode);
xQueryEvaluator.iterator();
```

The Java Convert method simply uses Rumble API to call serializeToJSONiq method we have previously created making the conversion complete. There is a small remark to take into consideration here. The test-cases that have error codes will cause a RumbleException. Due to time constraints we were unable to make a better classification and check for which error codes we actually need to perform XQuery to JSONiq conversion of test tag. Thus, those test-cases will not be converted and return same result instead. Check code below:

```java
private String Convert(String testString){
    try {
        return rumbleInstance.serializeToJSONiq(testString);
    } catch (RumbleException re) {
        return testString;
    } catch (Exception e){
        return testString;
    }
}
```

At the moment, Test Converter is dependent on Rumble as it was used to perform clean XQuery to JSONiq transformation. However, once Rumble is mature implementation supporting everything we have documented as missing, we will have an once and for all generated JSONiq Test Suite as the output of our work.

On the other hand, Test Driver will not change much. We have a similar architecture as we had with hard-coded conversion before but in a much cleaner approach. Test Driver can be used directly with either JSONiq or XQuery Test Suite depending on configuration which we will explain in Section 5.1. It will be helpful in improving the Rumble implementation by fixing and enable opening/closing issues and tracking the development progress of Rumble. Fixing the bugs and adding features will lead also consequentially lead to Item 2 list to drop down to 0.

Chapter 5

Conclusion and Future Work

As mentioned in Chapter 1, the high level idea of this work is to implement a Test Driver that can directly use QT3TS in order to test and verify Rumble implementation. However, during our work, we have managed to go beyond this scope and achieve even more which we will present in Section 5.1.

5.1 Result Summary

Apart from implementation of Test Driver, the results can be divided in total of 4 major areas:

1. Implementation of Test Driver for Rumble

2. Improvement of Rumble implementation

3. XQuery Parser extension of Rumble

4. Standalone JSONiq Test Suite

5.1.1 Implementation of Test Driver for Rumble

As we said, the published QT3TS Repository [W3C20] can be used to test any XML (XQuery, XPath, XSLT) implementation. Rumble as engine uses JSONiq language that inherits 95% of its features from XQuery. For each implementation, Test Driver has to be written in order to be able to use the QT3TS. And we have achieved that, we have fully operational Test Driver that can parse the QT3TS and execute it on top of Rumble. Depending on configuration it can be used in 3 modes:

1. Preferred way of testing Rumble implementation is using the original QT3TS and perform hard-coded conversion to JSONiq within Test Driver.

2. Verify implementation of XQuery Parser of Rumble by using the original QT3TS without performing hard-coded conversion to JSONiq within Test Driver.

3. Future way of testing Rumble implementation is using the JSONiq Test Suite generated from original QT3TS by Test Converter without performing hard-coded conversion to JSONiq within Test Driver.

5.1.2 Improvement of Rumble implementation

In this subsection we will discuss the impact and the usage that Test Driver had on Rumble implementation. Let us take another look at Table 3.2. This is one of the versions in which Test Driver itself was not very stable as it still had bugs in implementation. The fully stable version after which we had code freeze on Test Driver was implemented on 12th January 2021. After this version, we have performed manual inspection of failed and crashed test cases. We have opened over 50 issues on Rumble Repository [IFM+21]. Here, we will omit the test-cases that were skipped and aggregate categories in order to present simple classification as we had in Table 3.1.

In Table 5.1 we can see how did Rumble engine improved over period of 2 months by implementing bugfixes of the 50 above mentioned issues.

Scenario	Total not skipped test-cases		% of not skipped test-cases	
Date	12.01.21	12.03.21	12.01.21	12.03.21
Success	8837	9764	58.4	64.5
Fail	1351	999	8.9	7.6
Crash	4938	4372	32.7	28.9

Table 5.1: Rumble Implementation Improvement

The Crashes that are still visible in Table 5.1 can be improved by further classification of test-sets and test-cases into Item 1 or Item 2 that should be skipped as explained in Section 4.3. It can also be improved by extending Test Driver to support assert-xml and assert-serialization-matches.

5.1.3 XQuery Parser extension of Rumble

So far, Rumble was able to use only JSONiq as querying language. In order to convert the QT3TS - XQuery Test Suite to JSONiq, we decided to reuse the JSONiq Expression Tree already existing in Rumble. We first implemented XQuery Parser and XQueryTranslationVisitor that enabled us to obtain the JSONiq Expression Tree from query written in XQuery. We then implemented serialization that takes the JSONiq Expression Tree and outputs query written in JSONiq. The byproduct of such an exercise, resulted in extending the Rumble such that it is now able to operate using XQuery language as well.

5.1.4 Standalone JSONiq Test Suite

One of the biggest achievements of our work is producing purely JSONiq Test Suite similar to QT3TS for XQuery. This Test Suite uses .xml file format similar to QT3TS and it can be published and used by anyone to verify their JSONiq implementations. Anyone could write their own Test Driver and use our JSONiq Test Suite in similar fashion as we used the QT3TS one.

5.2 Future Work

Throughout this work, we have discussed many ideas and developed many prototypes. The prototypes are fully operational, however they can be extended or improved. The open problems that remained unresolved are:

- Test Driver and Test Converter - Extending Test Driver and Test Converter to support the last two missing assertions: serialization-matches and assert-xml.

- Test Driver - Implement a separate class for outputting the results. Right now, all the outputs are in form of log files in .txt or .csv file format. The step forward would be to implement class that would form a HTML web-page similar to one QT3TS has.

- Test Driver - extend Test Driver such that it can automatically detect bugs based on test cases that are not succeeding and automatically open/close issues on Rumble Repository.

- Test Converter - Improve serialization to JSONiq such that it has better file formatting with new lines and brackets.

- Test Converter - Some test cases are written in a way that they do not parse or cause other errors. Some of them should not be converted and Test Converter can be extended to not convert certain test cases based on their expected error code.

- Rumble - Improve XQuery Parser such that it passes more test cases using the original QT3TS.

- Rumble - Enable Rumble to automatically detect the under-laying query language and use the appropriate JSONiq or XQuery parser in order to execute the query.

Bibliography

[Cik20]   Can Berker Cikis. Machine learning with jsoniq, 2020.

[Cor17]   IBM Corp. Ibm websphere datapower gateways release notes, 2017.

[CZ18]   Bill Chambers and Matei Zaharia. *Spark: The Definitive Guide Big Data Processing Made Simple*. O'Reilly Media, Inc., 1st edition, 2018.

[DG04]   Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, pages 137–150, San Francisco, CA, 2004.

[FF13]   D. Florescu and G. Fourny. Jsoniq: The history of a query language. *IEEE Internet Computing*, 17(5):86–90, 2013.

[Fou13]   Ghislain Fourny. Jsoniq the sql of nosql, 2013.

[Fou18]   Ghislain Fourny. Ethz big data lecture - 263-3010-00l, 2018.

[GGL03]   Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. *SIGOPS Oper. Syst. Rev.*, 37(5):29–43, October 2003.

[IFM+20a]   Stefan Irimescu, Ghislain Fourny, Ingo Müller, Dan-Ovidiu Graur, Can Berker Çıkış, Renato Marroquin, Falko Noé, Ioana Stefan, Andrea Rinaldi, and Gustavo Alonso. Rumble supported error codes, 2017-2020.

[IFM+20b]   Stefan Irimescu, Ghislain Fourny, Ingo Müller, Dan-Ovidiu Graur, Can Berker Çıkış, Renato Marroquin, Falko Noé, Ioana Stefan,

Andrea Rinaldi, and Gustavo Alonso. Rumble unsupported types, 2017-2020.

[IFM+21]   Stefan Irimescu, Ghislain Fourny, Ingo Müller, Dan-Ovidiu Graur, Stevan Mihajlovic, Mario Arduini, Can Berker Çıkış, Renato Marroquin, Falko Noé, Ioana Stefan, Andrea Rinaldi, and Gustavo Alonso. Rumble github repository, 2017-2021.

[Iri18]   Stefan Irimescu. Jsoniq on spark, 2018.

[JSO]   Introducing json.

[jso20]   jsoniq.org. Jsoniq - the json query language, 2011-2020.

[Kay16]   Michael H. Kay. Qt3 test suite result summary, 2016.

[Kay20]   Michael H. Kay. Saxon: The xslt and xquery processor, 2020.

[KKWZ15]   Holden Karau, Andy Konwinski, Patrick Wendell, and Matei Zaharia. *Learning Spark: Lightning-Fast Big Data Analytics.* O'Reilly Media, Inc., 1st edition, 2015.

[MF21]   Ingo Müller and Ghislain Fourny. Iris-hep topical meeting (4 may 2020) - rumble: Jsoniq (query language) on spark, 2021.

[MFI+20]   Ingo Müller, Ghislain Fourny, Stefan Irimescu, Can Berker Cikis, and Gustavo Alonso. Rumble: Data independence for large messy data sets, 2020.

[Mih20]   Stevan Mihajlovic. Rumble repository, 2020.

[PQ95]   T. J. Parr and R. W. Quong. Antlr: A predicated-¡i¿ll(k)¡/i¿ parser generator. *Softw. Pract. Exper.*, 25(7):789–810, July 1995.

[Reg]   Regression testing.

[SKRC10]   K. Shvachko, H. Kuang, S. Radia, and R. Chansler. The hadoop distributed file system. In *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10, 2010.

[VMD+13]   Vinod Kumar Vavilapalli, Arun C. Murthy, Chris Douglas, Sharad Agarwal, Mahadev Konar, Robert Evans, Thomas Graves, Jason Lowe, Hitesh Shah, Siddharth Seth, Bikas Saha, Carlo Curino, Owen O'Malley, Sanjay Radia, Benjamin Reed, and Eric Baldeschwieler. Apache hadoop yarn: Yet another resource negotiator. In *Proceedings of the 4th Annual Symposium on Cloud Computing*, SOCC '13, New York, NY, USA, 2013. Association for Computing Machinery.

[W3C11]   W3C. Xquery/xpath/xslt 3.* test suite cvs repository, 2011.

[W3C13]   W3C. Xquery/xpath/xslt 3.* test suite, 1994-2013.

[W3C17]   W3C. Xquery 3.1: An xml query language w3c recommendation, 2017.

[W3C20]   W3C. Xquery/xpath/xslt 3.* test suite github repository, 2020.

[Whi15]   Tom White. *Hadoop: The Definitive Guide.* O'Reilly Media, Inc., 4th edition, 2015.

[Xid]   Xidel.

[Xqu]   An antlr4 implementation of xqdoc for xquery.

[ZCD⁺12] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 15–28, San Jose, CA, April 2012. USENIX Association.

[Zor13]   Zorba. Zorba nosql engine, 2013.

[ZXW⁺16] Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, October 2016.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

___

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

A Test Suite for Rumble

Authored by (in block letters):
For papers written by groups the names of all authors are required.

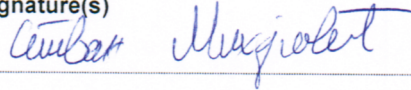| **Name(s):**	**First name(s):**
Mihajlovic	Stevan

With my signature I confirm that
 - I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
 - I have documented all methods, data and processes truthfully.
 - I have not manipulated any data.
 - I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date

Zürich, October 1, 2020

Signature(s)

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.