Evolving NoSQL Databases Without Downtime

Karla Saur
Intel Labs *
kara.saur@intel.com

Tudor Dumitras
University of Maryland, College Park
tdumitra@umiacs.umd.edu

Michael Hicks
University of Maryland, College Park
mwh@cs.umd.edu

Abstract

NoSQL databases like Redis, Cassandra, and MongoDB are increasingly popular because they are flexible, lightweight, and easy to work with. Applications that use these databases will evolve over time, sometimes necessitating (or preferring) a change to the format or organization of the data. The problem we address in this paper is: How can we support the evolution of high-availability applications and their NoSQL data online, without excessive delays or interruptions, even in the presence of backward-incompatible data format changes?

We present KVolve, an extension to the popular Redis NoSQL database, as a solution to this problem. KVolve permits a developer to submit an upgrade specification that defines how to transform existing data to the newest version. This transformation is applied lazily as applications interact with the database, thus avoiding long pause times. We demonstrate that KVolve is expressive enough to support substantial practical updates, including format changes to RedisFS, a Redis-backed file system, while imposing essentially no overhead in general use and minimal pause times during updates.

1. Introduction

NoSQL databases, such as Redis [36], Cassandra [4], and MongoDB [30], are increasingly the go-to choice for storing persistent data, dominating traditional SQL-based database management systems [6, 19]. NoSQL databases are often organized as key-value stores, in that they provide a simple key-based lookup and update service (i.e., “no SQL”). While these databases do not provide a formal language for specifying a schema, applications attach meaning to the format of the keys and values stored in the database. Keys are typically structured strings, and values store objects represented according to various formats [29], e.g., as Protocol Buffers [18], Thrift [5], Avro [3], or JSON [22] objects.

Database schemas change frequently when applications must support new features and business needs; for example, multiple schema changes are applied every week to Google’s AdWords database [34]. Applications that use NoSQL databases also evolve data formats over time, and may require modifying objects to add or delete fields, splitting objects so they are mapped to by multiple keys rather than a single key, renaming of keys or value fields, etc. When changes are not compatible with the old version of an application, a straightforward way to deploy them in the field would be to shut down the running applications; migrate each affected object in the database from the old format to the new format; and then start the new versions of the applications.

Performing these changes online is more challenging. The parsers for Thrift, Protocol Buffers, and Avro provide some support for format changes e.g. by skipping unknown fields or by attempting to translate data from the writer schema into the reader schema [25]. This leaves the task of updating each object in the database (e.g., by iterating over all of its keys [40]). For large amounts of data, this can create an unacceptably long pause. As an extreme example, Wikipedia was locked for editing during the upgrade to MediaWiki 1.5, and the schema was converted to the new version in about 22 hours [49]. Developers could avoid shutting down the application by making the new format backward-compatible with the old format, but this would impose a significant constraint on the future evolution of the application. It may also be possible to grant applications read-only access to the old database while the migration takes place, but applications that have even occasional writes will suffer.

One way to avoid downtime and database locking is to migrate data lazily. When the updated application accesses an object in the old format, the object is converted to the new format on-the-fly. Thus, the long pause due to migrating the data is now amortized over the updated application’s execution, causing slower queries immediately after the update but no full stoppage. Currently, the task of implementing lazy data migration falls on the developer: applications are written to expect data in both old and new formats and
to migrate from to the new format when the old one is encountered [7, 37, 39, 45]. This approach adds work for the programmer and results in code that mixes application and format-maintenance logic. Since there is no guarantee that all data will ultimately be migrated, the migration code expands with each format change, becoming more confusing and harder to maintain.

This paper presents KVolve¹, an extension to the popular Redis NoSQL database, as a solution to the problem of evolving a NoSQL database online. To simplify development, KVolve presents the logical view to applications that data is at the newest version of the format. Rather than convert all data at once, keys and values are converted as they are accessed in a way that requires almost no changes to applications—they simply indicate the data version they expect when they connect to the database, and they are permitted to proceed if their expected version and the logical view’s version match. When a data upgrade is installed, applications with an incompatible version must update themselves. They can do this either by simple stop-and-restart (to the new application version), or they can use dynamic software updating (DSU) [9, 16, 21, 33] or concurrent application switching, as in parallel AppEngine [17], to avoid lost application state and/or shorten pause times.

KVolve piggybacks on client commands to trigger data conversions automatically. To track its progress, KVolve attaches a version identifier to the value of each entry, converting only those keys/values that are out of date. Conversions are written by the developer, and are specified as functions from the old key/value format to the new one. KVolve does not support conversions in which a new key/value is a function of several old key/values; this is necessary to maintain the required logical view. Many applications we have considered easily satisfy this restriction.

Experiments with our implementation show that KVolve imposes essentially no overhead during normal operation. Using the standard Redis benchmark that repeatedly queries the database (the worst case as far as application performance is concerned), we observe the overhead to be in the noise. We also find that laziness significantly reduces the update-time disruption. We applied KVolve to upgrade redisfs [24], a full-featured file system that uses Redis as the back end. The upgrade in question (to version 0.7), required a significant data format change: keys representing file directories were renamed, and data containing the file contents were compressed. While KVolve took care of the changes to the database, we used Kitsune [21], a whole-program updating framework for C, to dynamically update the redisfs driver. Doing so allowed us to seamlessly maintain the file system mount point and other in-memory state. The result was zero downtime. We also tested our approach on a social networks modeling tool called Amico [1], and showed that we could reduce the pause time for renaming all 792,711 keys from 100 seconds to almost zero seconds with KVolve.

In summary, we make three contributions:

- We define a general approach to updating data formats in NoSQL databases lazily. For large databases, this approach gives the appearance of an instant update, as the applications can start using the new data format without delay, and it amortizes the cost of migration over the normal execution. Additionally, this approach does not limit the client-side interactions (e.g., to read-only execution) and requires very few changes to applications. (Section 2.)

- Our implementation with the Redis NoSQL database requires minimal changes to Redis itself: it is implemented in a modular way and does not interfere with Redis command processing. Moreover, although our implementation extends Redis, the architecture we present can be applied to other NoSQL databases; we include a discussion on extrapolating KVolve principles to MongoDB [30]. (Section 3.)

- Experiments show that KVolve adds essentially no overhead during normal operation, with the corresponding benefit of very little performance degradation during the lazy update rather than a full stoppage, especially compared to the eager approach. We also show how KVolve can be integrated with the Kitsune DSU [21] to perform end-to-end upgrades with no downtime (Section 4.)

To the best of our knowledge (see Section 5), KVolve represents the first general-purpose solution to the problem of evolving a NoSQL database without downtime. We plan to make our code freely available.

2. Overview

This section presents an overview of KVolve, describing its various pieces and how they fit together using a simple example.

2.1 Background on Redis and NoSQL databases

NoSQL databases are distinguished from traditional relational database management systems (RDBMSs), in supporting a very simple, lightweight interface. Our focus is on a variant referred to as key-value stores which, as the name implies, focus on mapping keys to values. We chose to work with the key-value store Redis [36], one of the most widely used key-value databases [43]. The main operations in Redis, in particular, are: GET k, which returns the value v to which k maps in the database (or “none” if none is present); and SET k v, which adds (or overwrites) the mapping k → v in the database. Redis is single-threaded and preserves the in-memory data by serializing the data to disk periodically. (We discuss concurrency and fault-tolerance in Section 3.6.)

Redis supports variations of these operations (e.g., setting values only if no prior version exists, or defining mappings

¹ KVolve stands for Key-Value store evolution.
that will time out). It also supports additional data structures (the main data structures include strings, sets, lists, hashes, and sorted sets) and their corresponding operations, such as appending an item to a list. KVolve supports 36 primary Redis commands and their variations, including support for all of the main Redis data structures. Although we have not yet implemented all commands for KVolve, we expect this to be straightforward, as discussed in Section 3.2. We also speculate on implementing KVolve for another NoSQL database, MongoDB, in Section 3.6.

Many applications store string values that adhere to formats such as JSON [22], Avro [3], or Protocol Buffers (“Proto Bufs”) [18]. In the example in this section, we focus on Redis databases that store JSON objects. JSON defines four primitive types: numbers, strings, booleans, and null. It also defines two container types: arrays, which are an ordered list of values of the same JSON type; and objects, which are an unordered collection of values of any JSON type, with field labels. Our basic approach should apply to other formats as well, and it would arguably be easier to do so, as Avro and Proto Bufs also define a notion of schema that could be analyzed to understand the effect of a change.

A common programming practice for key-value stores is the use of namespaces, which conceptually divide up the kinds of objects stored in the database. Redis does not provide native support for namespaces, but rather advises their use by convention: Keys should have the format n: k where n identifies the namespace, and k identifies the proper key name. The assumption is that objects in namespace n will all have the same type. Of course, n might be further partitioned into sub-namespaces, as required. We also support having a global namespace, treating all of the keys in the database as the same type.

### 2.2 Example data format change

Consider the example of an online store (adapted from Sadalage and Fowler [39]). Such an application may store purchase orders in its database. Keys have the format order: n where order is the namespace, and n is a unique invoice number. Such keys map to values that describe a purchase, formatted as JSON as shown in Figure 1a.

Suppose we wish to upgrade the application to support differentiated pricing, which necessitates changing the data format as follows: rename the field price to fullprice, and insert a new field named discountedPrice that is a markdown from the original price. The updated orderItems array (the last element of the JSON object) for the example, that adheres to the new format, is shown in Figure 1b.

To support migrating from the old to the new version of the application, Sadalage and Fowler suggest that the programmer can modify the new version’s code to essentially handle both formats, and migrate data from the old format to the new one when it is encountered.

This approach is efficacious but creates extra effort for the application programmer, as they have to write code that deals with the differing versions, and this code must reside in the application indefinitely. Sadalage and Fowler also do not explicitly consider the complications of concurrent clients, which could read/upgrade the data at the same time and produce inconsistencies. KVolve avoids this problem by ensuring that the key will be updated exactly once and that subsequent client requests will only read up-to-date data.

### 2.3 Database upgrades using KVolve

KVolve is a minimal extension to standard Redis with some update-enabling elements. (The implementation inner workings are described in detail in Section 3.2.) KVolve’s archi-
tecture is depicted in Figure 2, with its update elements depicted in blue. In particular, the figure shows (at the top) Client\_U installing an update specification in the database, signaling that the database’s contents need to be upgraded. And it shows two clients that are interacting with Redis, Client\_X and Client\_Y, trying to retrieve key Kx and key Ky respectively. The connection order is as follows: Client\_X connects first at v0 and runs for a while. After the update is installed by Client\_U, Client\_Y connects at v1.

An update specification indicates the expected, current version identifier (in the figure, it is v0); the new version identifier (here, v1); and code that can be used to migrate data from the current version to the new version (not shown). This code is written in C with the help of some template functions, described in detail in Section 2.4.

When a client connects, it declares what data version(s) it expects for each namespace it will use. KVolve uses this information to ensure that all clients are connected at only the current version of the namespace’s schema, and does not allow clients to connect that request outdated versions. When a version is declared for a specific namespace, KVolve automatically adds the version number to values so that KVolve may track what data it must lazily migrate when an update is requested. The version information also allows KVolve to help orchestrate a switchover to the newest application version by notifying out-of-date clients when/if they fall behind the logical version of the data they are using.

The bottom of the figure shows Client\_Y interacting with the database after the update and illustrates a lazy update taking place. Each requested value is checked by KVolve to see if an upgrade is necessary. Now, just after an update, there will be no keys at the newest version—all of the data is still at the last (or even older) versions. As such, KVolve performs an update on the value valY before having Redis return the value to the client at v1 with the value valY{apid}. The newly updated version information and value is also stored in Redis, and the old version of the key is freed.

In the figure, we see that Client\_X is still at v0. When the update is submitted, KVolve closes the connection to all outdated clients, and they will not be permitted to reconnect with the old version of the software. If we consider our purchase order example, we can see that the new version of the data will not work for older clients: They will be expecting JSON objects to contain field price but the data will (eventually) be stored under fields discountedPrice and fullPrice, instead. As such, the old client cannot safely access the data any longer, and must be terminated.

When data is updated in a backward-incompatible manner, clients designed to use that data will have been re-coded to use the new format. As such, when the old version is notified, it can start a new-version replacement that connects to the database. (To continue, Client\_X must upgrade to v1.) The new version then migrates the data as it accesses it, as described above. This approach simplifies application development as each application version can assume a particular data version, and KVolve ensures data is migrated as it is needed. In fact, the only changes required to an application to support lazy data updates are to (a) change the call to connect to Redis to also declare the expected version of the data, and (b) to gracefully terminate and upgrade to the new version when the connection is closed. Graceful termination involves saving any final state locally if necessary, as database access is no longer allowed. These changes amount to only a few lines of code for the entire application.

2.4 Describing data updates

The programmer must write update functions that will convert the old version of a key and/or value to the new version. These functions will be applied on demand by KVolve, but we want to present the logical view to the application that the data was converted eagerly.

An example update function is given in Figure 3. The old key (a string) and value (binary data) are passed in by reference, and the function will update them to the new versions via these references. In this case, the body of the function uses the Jansson library [26] to implement the change to the purchase order example from Figure 1 described in Section 2.2; the last two lines update the value (the key is not changed). This function is installed by an initialization function called when the update is loaded in KVolve:

```
kvolve_upd_spec("\"order\",\"order\", 0, 1, 1, test_fun_updval );
```

This indicates that the order namespace doesn’t change, from version 0 to version 1, while the test_fun_updval should be called for each key in the namespace order.

Namespaces can be changed without requiring an update function. For example, in the Amico program described later in Section 4.3, the keys are renamed from the namespace prefix of amico:followers to the namespace prefix of amico: followers : default . To describe this update, the update writer should specify:

```
kvolve_upd_spec("\"amico: followers \",
\"amico: followers : default \", 1, 2, 0);
```

where the version numbers are 1 and 2, and the 0 indicates that there are no functions to manipulate the value.

In order to preserve the illusion that the updates are done eagerly (while they are actually done lazily), the new key or value should be a function of only the old key or value, and not any other data in the database, since there is no guarantee about the order this data will be changed. Therefore, the only restriction to updates that KVolve can do lazily is that a new key/value format cannot be the format of several old key/value formats. An update function may query the database from within the update function (such as querying a constant that will not change throughout the update), but the data change should not depend on other data that may change.
3. KVolve Implementation

We designed KVolve to be as modular as possible and to avoid disrupting Redis’s normal command-processing control flow. This section describes our implementation of KVolve as an extension to Redis.

3.1 Design Goals

When designing KVolve, we wanted it to be general-purpose, efficient, and easy to maintain. We initially built an application-side wrapper written in Python, which has the virtue that no changes to Redis are required at all. But this wrapper is limited to Python Redis clients, and imposes a fair bit of overhead (5–8% on normal operations). We then experimented with a network proxy, written in C, that intercepts Redis-application communications. This approach also requires no changes to Redis, and supports applications written in any language. But we still found it to be slower than we’d like (about 3% overhead), and tricky to write, especially when trying to ensure transaction atomicity.

In the end, we designed KVolve as a separate library that is compiled into Redis itself. Our changes add a \texttt{vers} version field to the Redis \texttt{robj} data structure, which represents data values stored in the database. They also hook the command processing logic to support lazy migration. In total, 7 lines of Redis (v2.8.17) needed to be changed: 2 lines of code to add and initialize \texttt{vers} to a default value, 2 lines of code to add a function call into the KVolve library right before Redis processes commands, and 3 lines of code to allow the \texttt{vers} data to be serialized and deserialized to and from disk using the normal Redis serialization commands. This approach has proved to be the most efficient, most general, and easiest to get right. It should also be easy to maintain: The data structure we modified to add a version field has not changed significantly for several years, in particular, between versions 2.0.0 (2010) - 3.0.2 (2015). We believe the basic KVolve design and architecture should work for other key-value stores; we discuss MongoDB in particular in Section 3.6.

KVolve currently supports 3 Redis commands and all of the main Redis data structures (string, set, list, hash, sorted set). The commands we implemented for our prototype were chosen after examining roughly 30 programs on GitHub that used Redis to see which Redis commands were most commonly used.

3.2 KVolve implementation overview

KVolve works by preprocessing commands coming in from the client before passing them along to Redis, as depicted in Figure 4. In Step 1, the client issues the command set \texttt{Kx my_val}. In Step 2 \texttt{kvProcessCmd}, KVolve’s hook, is called to preprocess the command (the dashed green box is the KVolve library). Once the KVolve preprocessing is complete (which might involve changes to data’s contents and version field), control returns to normal Redis. In Step 3, Redis’s \texttt{processCmd} function calls the function pointer shown in blue (which depends on the choice of command—here it is \texttt{procSet} because the client requested set), and this adds the affected object to the database, including any changes to the version field set during KVolve’s processing. Finally, in Step 4 Redis responds to the client’s request, acknowledging to the client that it successfully executed the set command.

All of this is sure to be atomic because Redis is single-threaded: it processes each command it receives in its entirety before moving to the next. Redis provides commands, such as \texttt{multi}, that can be used to execute a group of commands atomically; KVolve’s design works in concert with such commands. We also believe that KVolve’s basic “interceptor” architecture would work in a multi-threaded setting; we discuss this setting further in Section 3.6.

All of KVolve’s operations are informed by a data structure it maintains called the \texttt{version hash table}. This hash table is constructed from the update specifications submit-
Keyed vs. string types

Redis supports five main data structures. Strings are the only type of data structures that are not containers, in that the value consists of a structure storing a single string. For that reason, we discuss string types separately.

3.3 String types

Redis supports five main data structures. Strings are the only type of data structures that are not containers, in that the value consists of a structure storing a single string. For that reason, we discuss string types separately.

---

2 As described in Section 2.1, namespaces are commonly used in key-value stores and conceptually divide up the kinds of objects in the database.
deletes the old robj and replaces it with the new one from the parsed client’s request. Therefore, all that KVolve must do is set the most current version string in the robj for the namespace of the key. (Remember that there is no need to attempt to update the value in the key, because the client’s provided value is guaranteed to be at the up-to-date version.)

If this set occurs after a namespace change, KVolve must delete the old value for the key to ensure that deprecated key versions are not unnecessarily bloating Redis. For example, in a change to redisfs (presented in Section 4.2), an old namespace key was named skx:/ but after an update, the new namespace postfixes DIR such that the key is now named skx:DIR:/ . If the user were to set the key skx:DIR:/root before getting (and updating) it, this would leave the old key skx:/root still in the database. Therefore, KVolve must check to see if the existing version under the old namespace exists, and if it does, delete it. It does this by first checking if the namespace had any previous namespace changes. If not, it does nothing. If so, it checks and deletes the old key if necessary. At this point, KVolve returns control to Redis, and Redis adds the robj structure to the database, which also contains the updated version string to be retrieved later if necessary.

### 3.4 Sets, hashes, lists, and sorted sets

The other Redis data structures are containers of sub-values. Although the exact substructure is different for each type, the basic implementation is the same, so we describe them together. The base of Redis containers are all robj structures, and they store the actual data in other ways to maximize efficiency of both space and speed. Figure 5 shows an example in columns two and three of value robjs that contain a hash of strings and a set of integers respectively. Because not all sub-values (individual set members, individual list items, etc) are stored directly in a robj structure, it is not possible to store the version information for the individual sub-values without further modifying Redis. Therefore KVolve must store the version information in the robj of the set/list/etc container robj itself, and KVolve must keep all members of the sub-values at the same version.

**Getting elements from container types** The process for doing a get on one of the container elements is exactly the same as for the string type described in Section 3.3, except that if necessary, the update is performed for all of the sub-elements. KVolve uses a Redis-provided iterator to access and update the sub-elements as necessary. If a container type has a namespace change, the rename process is exactly the same as it is for string types, as the keys are the same regardless of value type.

**Setting container types** The set process for container types is essentially equivalent to the get process described in the previous paragraph. For example, suppose key foo maps to the set of strings \{ ‘s1’, ‘s2’, ‘s3’ \} at version 1. Then say an update to version 2 replaces the letter s with the letter t for each element in the set. If a command, post-update, wishes to add string ‘t4’ to foo’s set, then the version in the set’s robj must be updated, along with the other values in the same set. This is done by performing a get to update the sub-elements, changing the value for foo to be \{ ‘t1’, ‘t2’, ‘t3’, ‘t4’ \}. KVolve will also update the version in the robj to be 2, and then return control to Redis, which will add ‘t4’ to the set. This process is similar for lists, sorted sets, and hashes: when any sub-element is accessed, KVolve iterates over and updates all the other sub-elements.

### 3.5 Installing an update

We conclude with a few more details about how updates are installed in KVolve.

The programmer compiles the update specifications into a shared object file that she can direct KVolve to load. In particular, the shared object file will contain the update functions (e.g., Figure 3), and a function kvolve_declare_update, which consists of a sequence of calls to kvolve_upd_spec that define the per-namespace changes defined by the update (again, see Section 2.4 for examples).

A client (or system admin) can load an update into KVolve by issuing a (repurposed Redis) command that includes the path to the location of the shared object. KVolve will load the shared object from the indicated location, which results in the kvolve_declare_update being run (it is declared as a “constructor” for the shared object, so it will be automatically called immediately after the shared object is loaded). For each call to kvolve_upd_spec (...) that is made, KVolve verifies that the specified old namespace exists at the specified old version, and that the specified new version does not exist (which could happen if a different client already loaded the update). In the case of a namespace change, KVolve verifies that the specified new namespace does not exist. If all these checks pass, the code is registered and loaded into the version hash table, so that it will be applied to any keys matching the namespace and version.

At this point, KVolve will close the connection to all clients using the old version of the namespace(s). (Clients not using the updated namespace(s) will not be affected.) Redis provides a client id for each connection, and as mentioned these ids are stored in the version hash table as each client connects. Note that this does not automatically kill the client’s application, it just closes the connection and the client will not be allowed to reconnect until it is upgraded to the new version.

Note that our current implementation stores these updates indefinitely. We find that the update functions take up a small amount of space relative to the rest of the data. (If an entry is stale and never accessed, then there is no reason to spend the overhead of updating it.) However, if program updates are very large or very frequent, one option would be to have a background thread forcing updates to outdated data, and freeing the update information once all entries have been updated.
3.6 Discussion

We conclude with some further discussion about KVolve’s implementation and applicability to other NoSQL database implementations.

**Concurrency.** As mentioned at the start of this section, Redis is single-threaded, makes it simple to see that KVolve’s extra per-command processing (or multi-command processing, when using the multi command) for data migration is atomic. Other NoSQL database implementations are multi-threaded, but we believe the basic KVolve architecture should still apply. These systems employ some form of synchronization to ensure atomicity, so KVolve’s processing can piggyback on its use. That said, KVolve effectively turns read operations into write operations while data is being migrated, which will impact the use of reader-writer locks for concurrency control, and possibly necessitate the use of a different synchronization protocol.

**Durability.** KVolve naturally piggybacks on the durability support of the underlying database system. In the case of Redis, durability is achieved through periodic checkpoints, rather than per transaction, and failures are rolled back to the previous checkpoint. KVolve adds version information to Redis data, and this information is checkpointed with the rest of the data. In addition, all metadata about an update is atomically stored to the database and so will also be restored on a crash, allowing the data migration to continue from where it was at the last checkpointed. We thoroughly tested this mechanism while developing KVolve. We believe that KVolve would work for other durability mechanisms too, such as per-transaction journaling (as in MongoDB).

**Other Data Models.** Different NoSQL databases support different data models and commands, and an implementation of KVolve for one of these would have to be adapted accordingly. While such an adaptation would require substantial future work, we believe that the basic design and implementation principles should carry over. In the remainder of this section we spend some time discussing how KVolve might apply to MongoDB [30], another popular NoSQL database.

MongoDB stores data as `documents` encoded using BSON [10], a binary representation of JSON. To implement KVolve versioning in MongoDB, we could add a version field to class `Document` objects. To handle commands, we could intercept calls to command handler’s process function. Logic to support basic query processing (“reads”) would be along the same lines as for Redis. Writes to MongoDB BSON documents would be similar to data updates to JSON string values stored in Redis. MongoDB writes are atomic at the document level, so there would be no concurrency concerns about one client reading a partially-updated document (value). MongoDB supports `references` from one document to the next, which are similar to in-memory pointers in normal programs. Thus, references that might change their referents or invariants during update can be handled similarly to pointers in DSU systems, e.g., by maintaining a map between old and new referents [21]. MongoDB also supports `indexes`, which are datastructures that support faster lookups. To retain their efficiency, indexes should probably be migrated eagerly; i.e., the new index is constructed based on the old one in its entirety, rather than piecemeal. Finally, many NoSQL databases support distributed implementations, for scalability. These can be handled by ensuring updates are installed atomically across the cluster—installation (without migration) is a relatively fast operation, so the cost of synchronization should be reasonable, but future work is required to explore the costs and benefits of different implementation approaches.

### 4. Experimental results

This section considers the performance impact of KVolve, during normal operation and during an update. Our experimental results are summarized as follows:

- Using the standard benchmark that is included with Redis, we found that KVolve adds essentially no overhead during normal operation, and we determined that storing the version and update information in Redis adds only about a 15% overhead in space.
- We updated the redisfs file system which included renaming some keys and compressing some data stored in keys, and found the operating overhead to be in noise, and the pause time to be close to zero as opposed to 12 seconds for an offline migration of the data.
- We updated the Amico social network system and found no added overhead, with a pause time of close to zero as opposed to 87 seconds for an offline migration of the data.

All experiments were performed on a computer with 24 processors (Intel(R) Xeon(R) CPU E5-2430 0 @ 2.20GHz) and 32 GB RAM with GCC 4.4.7 on Red Hat Enterprise Linux Server release 6.5. All tests report the median of 11 trials, and communication was via localhost with ~.03 ms latency.

<table>
<thead>
<tr>
<th>Program</th>
<th>Max RSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redis, empty</td>
<td>7.7MB</td>
</tr>
<tr>
<td>Redis, 1M 10-byte values</td>
<td>112.1MB</td>
</tr>
<tr>
<td>KVolve, empty</td>
<td>7.7MB</td>
</tr>
<tr>
<td>KVolve, 5 namespaces, 1M 10-byte values</td>
<td>128.6MB</td>
</tr>
</tbody>
</table>
Table 1: Redis-bench with single instructions for Redis vs KVolve (times in seconds, median of 11 trials)

<table>
<thead>
<tr>
<th></th>
<th>Redis</th>
<th>No NS, KVolve</th>
<th>With NS, KVolve</th>
<th>&amp;Prev NS, KVolve</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>time siqr</td>
<td>time siqr</td>
<td>time siqr OH</td>
<td>time siqr OH</td>
</tr>
<tr>
<td>String Get</td>
<td>58.83 (0.25)</td>
<td>58.08 (0.11)</td>
<td>-0.77%</td>
<td>58.46 (0.26)</td>
</tr>
<tr>
<td>Set Pop</td>
<td>58.20 (0.10)</td>
<td>58.25 (0.20)</td>
<td>0.09%</td>
<td>57.72 (0.31)</td>
</tr>
<tr>
<td>List Pop</td>
<td>58.47 (0.41)</td>
<td>58.93 (0.68)</td>
<td>0.79%</td>
<td>58.71 (0.15)</td>
</tr>
<tr>
<td>String Set</td>
<td>63.52 (1.04)</td>
<td>63.66 (0.70)</td>
<td>0.22%</td>
<td>65.39 (1.49)</td>
</tr>
<tr>
<td>Set Add</td>
<td>61.52 (0.77)</td>
<td>61.11 (0.24)</td>
<td>-0.67%</td>
<td>61.25 (0.97)</td>
</tr>
<tr>
<td>List Push</td>
<td>59.49 (0.66)</td>
<td>59.55 (0.56)</td>
<td>0.10%</td>
<td>60.02 (0.74)</td>
</tr>
</tbody>
</table>

Table 2: Redis-bench with 10 pipelined instructions for Redis vs KVolve (times in seconds, median of 11 trials)

<table>
<thead>
<tr>
<th></th>
<th>Redis</th>
<th>No NS, KVolve</th>
<th>With NS, KVolve</th>
<th>&amp;Prev NS, KVolve</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>time siqr</td>
<td>time siqr</td>
<td>time siqr OH</td>
<td>time siqr OH</td>
</tr>
<tr>
<td>String Get</td>
<td>9.73 (0.30)</td>
<td>9.75 (0.27)</td>
<td>0.21%</td>
<td>9.96 (0.25)</td>
</tr>
<tr>
<td>Set Pop</td>
<td>9.88 (0.41)</td>
<td>9.52 (0.35)</td>
<td>-3.64%</td>
<td>10.04 (0.51)</td>
</tr>
<tr>
<td>List Pop</td>
<td>9.60 (0.32)</td>
<td>9.71 (0.25)</td>
<td>1.15%</td>
<td>9.55 (0.43)</td>
</tr>
<tr>
<td>String Set</td>
<td>13.77 (0.26)</td>
<td>14.56 (0.18)</td>
<td>5.74%</td>
<td>14.56 (0.32)</td>
</tr>
<tr>
<td>Set Add</td>
<td>13.97 (0.57)</td>
<td>14.09 (0.32)</td>
<td>0.86%</td>
<td>14.35 (0.71)</td>
</tr>
<tr>
<td>List Push</td>
<td>14.22 (0.39)</td>
<td>14.38 (0.25)</td>
<td>1.13%</td>
<td>14.40 (0.41)</td>
</tr>
</tbody>
</table>

1 request per round trip), and with a single key (getting or setting a single key multiple times). However, Redis-bench allows many different configurations. For a longer benchmark, we increased the number of operations to 5 million operations and for a more realistic benchmark we performed these operations over 1 million keys, leaving the rest of the default settings alone. We ran this experiment over localhost which had a latency of ∼0.03ms. We chose three types of get operations (string gets, set pops, and list pops) and three types of set operations (string sets, set adds, and list pushes), as these were part of the default benchmark operations test suite.

Table 1 shows the steady state overhead of this experiment. We show unmodified Redis in column 2 for comparison and broke the overhead into separate categories: KVolve with no namespace declared (causing KVolve to return immediately for each key) in column 3, KVolve with a single namespace declared (causing a hash lookup and a version check for each key) in column 4, and KVolve with a previous namespace declared but no previous keys at the old namespace (causing a hash lookup, a version check, and a string concatenation to look for a non-existent previous key) in column 5. Each sub-column of Table 1 shows the total time for the test, the siqr (Semi-Interquartile Range) to show the variance, and the overhead as a comparison against unmodified Redis. We ran this benchmark many times with various configurations (multiple namespaces, less or fewer keys, less or fewer clients, etc) and found that the overhead varied generally around ±3%, with no consistent pattern between any of the tests, even repeated tests with the exact same setup.

The numbers presented in the table show some negative and some positive overhead, reflecting this variation. Notice that the siqr numbers show that the variance is relatively high, as high as 1.49s for setting strings with KVolve and a namespace, shown in the sixth row of the fourth column.

Table 2 shows a modification of the original overhead experiment, using a pipeline to feed 10 instructions into each round trip to Redis-bench over localhost. This reduced the I/O overhead, putting more emphasis on KVolve operations. We found that these numbers showed a bit more overhead, and allowed us to bound the overhead at 5.74% for 10 subsequent pipelined instructions. This test demonstrated that although there is some overhead added by KVolve, for the non-pipelined version and most commonly-used scenario (Table 1), the overhead is mostly buried in I/O and very low overall. (In our test programs, described next, Amico pipelined at most 3 instructions per round trip, and redisfs did not use pipelining.)

In addition to time overhead, KVolve incurs some additional memory overhead due to tracking the version information. Table 3 shows the maximum resident set size as reported by pm. Empty, Redis and KVolve take up about the same amount of size in memory. With 1 million keys each mapping to 10-byte values and with 5 namespaces declared, KVolve takes up about 16.5MB (~15%) more memory than unmodified Redis. This includes the extra version field (4 bytes) on each value structure, the amount of space it takes to store the version lookup information and hash table, and any extra padding that may be automatically added to the additional structures.

2015/10/24
Redisfs [24] uses Redis as the backend to the FUSE [47] file system. The inode information, directory information, and all file system data are stored in Redis. On startup, FUSE mounts a directory with Redis as the backend, and a user can perform all of the normal operations of a file system, with the data silently being stored in Redis. Redisfs has 8 releases, ∼2.2K lines of C code each. In redisfs.5, released March 4th, 2011, file data is stored in a Redis as a binary string with no compression, and the directory keys have the format skx:/path/todir. In redisfs.7, released March 11th, 2011, file data is compressed using zlib, and directory keys have the format skx:DIR:/path/todir. (Note that redisfs.6 contained an error and was retracted, so we use versions .5 and .7.) This change makes it impossible to view the directories or any of the files using redisfs.7 for any files created using redisfs.5.

In all versions, the inode data is stored across about 12 Redis keys with meta information such as modification time, creation time, file size, and also the data itself. All file system information is represented in redisfs with four namespaces: the skx:/ namespace to represent the directories (which is updated to skx:DIR/ in redisfs.7), the skx:NODE namespace to represent the inodes (some of which is updated to add compression in redisfs.7), and also skx:PATH for paths to directories and skx:GLOBAL to track internal structure, neither of which are updated. To make redisfs compatible with KVolve, we added the same 6 lines of code in both versions which consisted of an additional call to Redis on start-up to declare that we would be using those 4 namespaces at either version .5 or .7, along with a few additional lines of error handling to make sure that the namespaces were properly set. No additional changes were made to redisfs for KVolve.

We performed an update from redisfs.5 to redisfs.7, both by migrating the keys offline (referred to as the Eager version), and with KVolve to automatically rename the directory keys as they are accessed and to add compression to the files as they are accessed. In addition to updating redisfs with KVolve, we also used Kitsune [21], whole-program update software for C, to allow us to also dynamically update redisfs along with the data so that the users experience no downtime; the switchover from .5 to .7 is completely seamless. Normally, killing redisfs.5 and restarting at redisfs.7 also causes the mount point to be unmounted then remounted (causing the user to have to switch back into the mounted directory after remount), but with Kitsune, the mount point is not disrupted during the switchover. We used the file system benchmark PostMark [23] to generate a workload for redisfs, creating an initial 10,000 files ranging from 4-1024 bytes in 250 subdirectories plus the root directory, for a total of 251 directories. We ran PostMark outside the root directory mount point, accessing the files via full path name to avoid having to change directories due to the restart for the Eager (non-KVolve/Kitsune) version.

Figure 6 shows the results of the redisfs experiment. After about 60 seconds, PostMark switched from creating the new files to reading from or appending to existing files. Both KVolve and the Eager version had a very similar average Queries Per Seconds (QPS), as shown by the solid and finely dashed lines which correspond to the left y-axis. At 80 seconds, we killed redisfs.5. For KVolve, we used Kitsune to dynamically update to redisfs.7 without pause, maintaining the mount point so that the benchmark never lost access to the files or the directory structure, and KVolve continued to process queries throughout the update. For the Eager version, we halted all traffic to Redis and migrated the data, performing the renames and compression as necessary. In this update, not all of the keys needed to be updated, only the 251 directory keys that needed to be renamed and the 10,000 data keys that needed to be compressed. However, the database contained 123,002 total keys, and the to-be-updated keys were searched for in the database, adding to the pause time. This offline update process took about 12 seconds, as shown in Table 4.

In addition to showing the QPS lines, the green widely-dashed line in Figure 6 shows the number of lazy updates per second for KVolve, corresponding to the right y-axis. Immediately after the update, this number burst to about 3,000 keys per second, and quickly trailed off as keys were lazily updated. KVolve renamed the 251 directory keys, updated the version tag on all 112,752 keys in the skx:INODE namespace, and added compression to the 10,000 keys in that namespace that contained file data.

Overall the impact on the update experienced by redisfs was minimal, as the QPS dipped only slightly right after the update before it quickly returned to full speed around the 120 second mark of the experiment. After the update, the overall QPS was slower for both KVolve and redisfs because the files must be compressed and decompressed as they were accessed.

4.3 Amico

Amico [1] maps relationships in the style of a social network, defining a set of users and the relationships between them. Amico provides an API that allows queries
over a data set of users: a user may be following or be followed by any number of other users. Amico is backed by Redis, has 10 versions created between 2012-2013, and is written in ~200 lines of Ruby code. Amico version 1.2.0, released Feb 22, 2012, stores these relationships in 5 different types of Redis keys with the following prefixes: amico: followers, amico: following, amico: blocked, amico: reciprocated, and amico: pending. In version 2.0.0, released Feb 28, 2012 (the next consecutive version after 1.2.0), the developers added the concept of a “scope” so that there could be different graphs stored in Redis with prefixes to keep them separate, such as “school” network and a “home” network. The default name for the scope is “default”, such that all of the keys are prefixed with amico: followers: default for example. This change makes databases created with Amico 1.2.0 incompatible with Amico 2.0.0. To make Amico work with KVolve, we only changed the same 4 lines of code in each version to declare the namespaces right after Amico connects to Redis.

For this experiment, we used the LiveJournal data set from the SNAP [27] library. The LiveJournal data set has 4,847,571 nodes and 68,993,773 directed edges defined by ordered node id numbers A follows B such as 186032 2345471, which we shuffled into two separate files for reading in a random order. To create a workload, we started two programs with calls to the Amico 1.2.0 API: one program to read from the first random file and add nodes to the Amico network, and one program to read from the second random file and perform queries over nodes in the network such as querying if USER A followed USER B or querying the number of followers of USER A. After letting the programs run for 900 seconds (15 minutes), the Redis database was filed with 792,711 keys containing nodes and edge data.

At the 900 second mark, as shown in Figure 7, we stopped both of the Amico 1.2.0 programs. For the Eager case (finely dashed line), we then updated all 792,711 keys, renaming them to have the default scope prefix in all of the key names. This migration took about 87 seconds as shown in Table 4. In addition to the pause, the Eager case shows a continued disruption until around the 1,000 second mark. After the migration was complete, we started the same writer/reader programs, this time using the Amico 2.0.0 API. For the KVolve case (solid line), we immediately started the two Amico 2.0.0 programs after the update so that the keys could be lazily migrated. Right at the update point, there is a ~2K drop in the QPS, before a brief spike and a return to the original rate. The widely-dashed green line corresponds to the right y-axis and shows the number of lazy updates that take place each second. Because this is a very large data set, many of the keys are not accessed immediately, taking full advantage of laziness. Although the lazy updates continue at a rate of about 500 per second at the 1,100 second mark, this does not significantly impact overall queries per second, as shown by the solid line maintaining a similar QPS before and after the update.

5. Related Work

In the realm of relational databases, the evolution of an application’s schema is characterized by the changes to the CREATE TABLE statements used to instantiate the schema in subsequent versions of the application. In practice, complex schema changes often require taking the application offline or locking the database tables, such as the update to Wikipedia that held a write lock for 22 hours [49]. Prior research has proposed supporting non-blocking schema changes by accepting out of date copies of database objects [46] or by implementing changes on-the-fly using triggers [38] or log redo [28]. Additionally, several professional tools can perform ALTER TABLE operations in a non-blocking manner [8, 31, 32, 42, 44]. Because these tools focus only on the database, the changes implemented must be backward compatible to avoid breaking the application logic. To avoid this limitation, the Imago system [14] proposed installing the new version in a parallel universe, with dedicated application servers and databases, which allowed it to perform an end-to-end upgrade atomically. This can be achieved in practice by deploying parallel AppEngine [17] applications, at multiple versions. However, this approach duplicates resources and exposes the new version to the live workload only after the data migration was completed.

In contrast, the F1 database from Google implemented an asynchronous protocol [34] for adding and removing tables, columns and indexes, which allows the servers in a distributed database system to access and update all the data during a schema change and to transition to the new schema at different times. This is achieved by having stateless database servers with temporal schema leases, by iden-
tifying which schema-change operations may cause inconsistencies, and by breaking these into a sequence of schema changes that preserve database consistency as long as servers are no more than one schema version behind. Google’s Spanner distributed key-value store [11] (which provides F1’s backend) supports changes to the format of keys and values by registering schema-change transactions at a specific time in the future and by utilizing globally synchronized clocks to coordinate reads and writes with these transactions. These systems do not address changes to the format of Protocol Buffers stored in the F1 columns or Spanner values [39] or inconsistencies that may be caused by interactions with (stateful) clients using different schemas [15].

Schema evolution in NoSQL databases is less well understood, as these databases do not provide a data definition language for specifying the schema. However, many applications attach meaning to the format of the keys and values stored in the database, and these formats may evolve over time. In particular, the values often correspond to data structures serialized using JSON [22] or a binary format like Thrift [5], Protobufs [18], or Avro [3]. The latter formats have schema-aware parsers, which include some support for schema changes, e.g. by skipping unknown fields or by attempting to translate data from the writer schema into the reader schema [25]. However, orchestrating the actual changes to the data and the application logic is entirely up to the programmer.

One approach to defining schema changes defines a declarative schema evolution language for NoSQL databases [41]. This language allows specifying more comprehensive schema changes and enables the automatic generation of database queries for migrating eagerly to the new schema. (While the paper also mentions the possibility of performing the migration in a lazy manner, which is needed for avoiding downtime, design and implementation details are not provided.) Another approach uses a DSL for describing data schema migrations for Haskell datatypes [20]. Many other approaches [2, 12, 35, 48] have focused on the problem of synthesizing the transformation code to migrate from one schema version to the next, and the transformation is then typically applied offline, rather than incrementally online. In this paper, we are not focusing on the problem of synthesizing the transformation code for now, and in any case it is a much simpler affair in the key-value setting. Rather, we focus on how to apply a transformation without halting service.

In practice, developers are often advised to handle all the necessary schema changes in custom code, added to the application logic that may modify the data in the database [7, 37, 39, 45]. This approach adds burden on the programmers, results in complex code that mixes application and schema-maintenance logic, does not provide a mechanism for reasoning about the correctness of schema changes performed concurrently with the live workload and often leave outdated entries in the database.

Our work is also related to the body of research on dynamic software updates [9, 16, 21, 33], which aim to modify a running program on-the-fly, without causing downtime. However, with the exception of a position paper [13], these approaches focus on changes to code and data structures loaded in memory, rather than changes to the formats of persistent data stored in a database.

6. Conclusions and Future Work

This paper has presented KVolve, a general approach to evolving a NoSQL database without downtime. KVolve adapts Redis to migrate data as it is accessed, reducing downtime that would otherwise result during a data upgrade, and minimizing required changes to applications. We find that KVolve imposes essentially no overhead when not performing an update, and minimal overhead when performing an update.

In the future, we would like to expand KVolve to work with Redis Cluster, a distributed implementation of Redis. We also would like to add direct support for programmer-specified, backward-compatible updates, which would support continued operation without restarting clients. Finally, we would like to streamline writing the transformation function with a domain-specific language, simplifying the update planning process.

We plan to release our code and make it freely available.

References

[1] Agora Games. Relationships (e.g. friendships) backed by redis. https://github.com/agoragames/amic0.


[45] stackoverflow.com. Are there any tools for schema migration for nosql databases?


