Lifetime-based memory management for distributed data processing systems

Lu, Lu; Shi, Xuanhua; Zhou, Yongluan; Zhang, Xiong; Jin, Hai; Pei, Cheng; He, Ligang; Geng, Yuanzhen

Published in:
Proceedings of the 42nd International Conference on Very Large Data Bases, New Delhi, India

DOI:
10.14778/2994509.2994513

Publication date:
2016

Document Version
Early version, also known as pre-print

Citation for published version (APA):
ABSTRACT

In-memory caching of intermediate data and eager combination of data in shuffle buffers have been shown to be very effective in minimizing the re-computation and I/O cost in distributed data processing systems like Spark and Flink. However, it has also been widely reported that these techniques would create a large amount of long-living data objects in the heap, which may quickly saturate the garbage collector, especially when handling a large dataset, and hence would limit the scalability of the system. To eliminate this problem, we propose a lifetime-based memory management framework, which, by automatically analyzing the user-defined functions and data types, obtains the expected lifetime of the data objects, and then allocates and releases memory space accordingly to minimize the garbage collection overhead. In particular, we present Deca, a concrete implementation of our proposal on top of Spark, which transparently decomposes and groups objects with similar lifetimes into byte arrays and releases their space altogether when their lifetimes come to an end. An extensive experimental study using both synthetic and real datasets shows that, in comparing to Spark, Deca is able to 1) reduce the garbage collection time by up to 99.9%, 2) to achieve up to 22.7x speed up in terms of execution time in cases without data spilling and 41.6x speedup in cases with data spilling, and 3) to consume up to 46.6% less memory.

1. INTRODUCTION

Distributed data processing systems, such as Spark [38], process huge volumes of data in a scale-out fashion. Unlike traditional database systems using declarative query languages and relational (or multidimensional) data models, these systems allow users to implement application logics through User Defined Functions (UDFs) and User Defined Types (UDTs) using high-level imperative languages (such as Java, Scala and C# etc.), which can then be automatically parallelized onto a large-scale cluster.

Existing researches in these systems mostly focus on scalability and fault-tolerance issues in a distributed environment [36][37][19][10]. However some recent studies [11][25] suggest that the execution efficiency of individual tasks in these systems is low. A major reason is that both the execution frameworks and user programs of these systems are implemented using high-level imperative languages running in managed runtime platforms (such as JVM, .NET CLR, etc.). These managed runtime platforms commonly have built-in automatic memory management, which brings significant memory and CPU overheads. For example, the modern tracing-based garbage collectors (GC) may consume a large amount of CPU cycles to trace living objects in the heap [29][15][4].

Furthermore, to improve the performance of multi-stage and iterative computations, recently developed systems support caching of intermediate data in the main memory [29][32][38][39] and exploit eager combining and aggregating of data in the shuffling phases [22][31]. These techniques would generate massive long-living data objects in the heap, which usually stay in the memory for a significant portion of the job execution time. However, the unnecessary continuous tracing and marking of such large amount of long-living objects by the GC would consume significant CPU cycles.

In this paper, we argue that distributed data processing systems like Spark, should employ a lifetime-based memory manager, which allocates and releases memory according to the lifetimes of the data objects rather than relying on a conventional tracing-based GC. To verify this concept, we present Deca, an automatic Spark optimizer, which adopts a lifetime-based memory management scheme for efficiently reclaiming memory space. Deca automatically analyzes the lifetimes of objects in different data containers in Spark, such as UDF variables, cached data blocks and shuffle buffers, and then transparently decomposes and stores a massive number of objects with similar lifetimes into a few number of byte arrays. In this way, the massive objects essentially bypass the continuous tracing of the GC and their space can be released by the destruction of the byte arrays.

Last but not the least, Deca automatically transforms the user programs so that the new memory layout is transparent to the users. By using the aforementioned techniques, Deca significantly optimizes the efficiency of Spark’s memory management and at the same time keeps the generality and expressibility provided in Spark’s programming model. In summary, the main contributions of this paper include:

• We propose a lifetime-based memory management scheme for distributed data processing systems and im-
implement a prototype on top of Spark, which is able to minimize the GC overhead and eliminate the memory bloat problems in Spark.

- We design a method that changes the in-memory representation of the object graph of each data item by discarding all the reference values. The raw data of the fields of primitive types in the object graph will be compactly stored as a byte sequence.
- We propose techniques to group byte sequences of data items with the same lifetime into a few byte arrays, thereby simplifying space reclamation. Deca automatically validates the memory safety of data accessing based on analysis of memory usage of UDT objects.
- We conduct extensive evaluation on various Spark programs using both synthetic and real datasets. The experimental results demonstrate the superiority of our approach by comparing with existing methods.

2. OVERVIEW OF DECA

2.1 Java GC

In typical JVM implementations, a garbage collector (GC) attempts to reclaim memory occupied by objects that will no longer be used. A tracing GC traces which objects are reachable by a sequence of references from some root objects. The unreachable ones, which are called garbages, can be reclaimed. Oracle’s Hotspot JVM implements three GC algorithms. The default Parallel Scavenge (PS) algorithm suspends the application and spawns several parallel GC threads to achieve high throughput. The other two algorithms, namely Concurrent Mark-Sweep (CMS) and Garbage-First (G1), attempt to reduce GC latency by spawning concurrent GC threads that run simultaneously with the application thread.

All the above collectors segregate objects into multiple generations: young generation containing recently-allocated objects, old generation containing older objects, and permanent generation containing class metadata. Based on the assumption that most objects would soon become garbages, a minor GC, which only attempts to reclaim garbages in the young generation, can be run to reclaim enough memory space. However, if there are too many old objects, then a full (or major) GC would be run to reclaim space occupied by the old objects. Usually, a full GC is much more expensive than a minor GC.

2.2 Motivating Example

A major concept of Spark is Resilient Distributed Dataset (RDD), which is a fault-tolerant dataset that can be processed in parallel by a set of UDF operations.

We use Logistic Regression (LR) as an example to motivate and illustrate the optimization techniques adopted in Deca. It is a classifier that attempts to find an optimal hyperplane that separates the data points in a multi-dimensional feature space into two sets. Figure 1 shows the code of LR in Spark. The raw dataset is a text file with each line containing one data point. Hence the first UDF is a map function, which extracts the data points and store them into a set of DenseVector objects and (lines 12–16). An additional LabeledPoint object is created for each data point to package its feature vector and label value together.

def this(data: Array[V]) = 
  this(data, 0, 1, data.length)
...

2.3 Life-time based Memory Management

We implement the prototype of Deca based on Spark. In Deca, objects are stored in three types of data containers:
UDF variables, cached RDDs and Shuffle buffers. In each data container, Deca allocates a number of fixed-sized byte arrays. By using the points-to analysis \cite{21}, we map the UDT objects with their appropriate containers. UDT objects are then stored in the byte arrays after eliminating the unnecessary object headers and object references. This compact layout would not only minimize the memory consumption of data objects but also dramatically reduce the overhead of GC, because GC only needs to trace a few byte arrays instead of a huge number of UDT objects. One can see that the size of each byte array should not be too small or too large, otherwise it would incur high GC overheads or large unused memory spaces.

As an example, the LabeledPoint objects in the LR program can be transformed into byte arrays as shown in Figure 2. Here, all the reference variables (in orange color, such as \texttt{features} and \texttt{data}), as well as the headers of all the objects are eliminated. All the cached LabeledPoint objects are stored into byte arrays.

The challenge of employing such a compact layout is that the space allocated to each object is fixed. Therefore, we have to ensure that the size of an object would not exceed its allocated space during execution so that it will not damage the data layout. This is easy for some types of fields, such as primitive types, but less obvious for others. Code analysis is necessary to identify the change patterns of the objects’ sizes. Such an analysis may have a global scope. For example, a global code analysis may identify that the \texttt{features} arrays of all the LabeledPoint objects (created in line 14 in Figure 1) actually have the same fixed size $D$, which is a global constant. Furthermore, the \texttt{features} field of a LabeledPoint object is only assigned in the LabeledPoint constructor. Therefore, all the LabeledPoint objects actually have the same fixed size. Another interesting pattern is that in Spark applications, objects in cached RDDs or shuffle buffers are often generated sequentially and their sizes will not be changed once they are completely generated. Identifying such useful patterns by a sophisticated code analysis is necessary to ensure the safety of decomposing UDT objects and storing them compactly into byte arrays.

As mentioned earlier, during the execution of programs in a system like Spark, the lifetimes of data containers created by the framework can be pre-determined explicitly. For example, the lifetimes of objects in a cached RDD is determined by the invocations of \texttt{cache()} and \texttt{unpersist()} in the program. Recall that the UDT objects stored in the compact byte arrays would bypass the GC. We put the UDT objects with the same lifetime into the same container. For example, the cached LabeledPoint objects in LR have the same lifetime, so they are stored in the same container. When a container’s lifetime comes to an end, we simply release all the references of the byte arrays in the container, then the GC can reclaim the whole space occupied by the massive amount of objects.

Lastly, Deca modifies the application code by replacing the codes of object creation, field access and UDT methods with new codes that directly write and read the byte arrays.

3. UDT CLASSIFICATION ANALYSIS

3.1 Data-size and Size-type of Objects

To allocate enough memory space for objects, we have to estimate the object sizes and their change patterns during runtime. Due to the complexity of object models, to accurately estimate the size of a UDT, we have to dynamically traverse the runtime object reference graph of each target object and compute the total memory consumption. Such a dynamic analysis is too costly at runtime, especially with a large number of objects. Therefore we opt for static analysis which only uses static object reference graphs and would not incur any runtime overhead. We define the \texttt{data-size} of an object to be the sum of the sizes of the primitive-type fields in its static object reference graph. An object’s \texttt{data-size} is only an upper bound of the actual memory consumption of its raw data, if one considers the cases with object sharing.

To see if UDT objects can be safely decomposed into byte sequences, we should examine how their data-sizes change during runtime. There are two types of UDTs that can meet the safety requirement: 1) the data-sizes of all the instances of the UDT are identical and do not change during runtime; or 2) the data-sizes of all the instances of the UDT do not change during runtime. We call these two kinds of UDTs as Static Fixed-Sized Type (SFST) and Runtime Fixed-Sized Type (RFST) respectively.

In addition, we call UDTs that have type-dependency cycles in their type definition graphs as Recursively-Defined Type. Even without object sharing, the instances of these types can have reference cycles in their object graphs. Therefore, they cannot be safely decomposed. Furthermore, any UDT that does not belong to any of the aforementioned types is called a Variable-Sized Type (VST). Once a VST object is constructed, its data-size may change due to field assignments and method invocations during runtime.

The objective of the UDT classification analysis is to generate the \texttt{Size-Type} of each target UDT according to the above definitions. As demonstrated in Figure 2, Deca decomposes a set of objects into primitive values and store them contiguously into compact byte sequences in a byte array. A safe decomposition requires that the original UDT objects are either of an SFST or an RFST. Otherwise the operations that expand the byte sequences occupied by an object may overwrite the data of the subsequent objects in the same byte array. Furthermore, as we will discuss later, an SFST can be safely decomposed in more cases than an RFST. On the other hand, objects that do not belong to an SFST or an RFST will not be decomposed into byte sequences in Deca. Apparently, to maximize the effect of our approach, we should avoid overestimating the variability of the data-size of the UDTs, which is the design goal of our following algorithms.

3.2 Local Classification Analysis

The local classification algorithm analyzes an UDT by recursively traversing its type dependency graph. For example, Figure 3 illustrates the type dependency graph of LabeledPoint. The size-type of LabeledPoint can be determined based on the size-type of each of its fields.

Algorithm 1 shows the procedure of the local classification analysis. The input of the algorithm is an annotated type that contains the information of fields and methods of the target UDT. Because the objects referenced by a field can be of any subtype of its declared type, we use a type-set to store all the possible runtime types of each field. The type-set of each field is obtained in a pre-processing phase of Deca by using the points-to analysis \cite{21} (see Section 5).

In lines 1–2, the algorithm first determines whether the
Algorithm 1: Local classification analysis.

Input: The top-level annotated type $T$;
Output: The size-type of $T$;
1 build the type dependency graph $G$ for $T$;
2 if $G$ contains the circle path then return RecurDef;
3 else return AnalyzeType($T$);
4 Function AnalyzeType($t_{arg}$);
5 if $t_{arg}$ is a primitive type then return StaticFixed;
6 else if $t_{arg}$ is an array type then
7 \hspace{1cm} $f_e$ ← array element field of $t_{arg}$;
8 \hspace{1cm} if AnalyzeField($f_e$) = StaticFixed then
9 \hspace{2cm} return RuntimeFixed;
10 else return Variable;
11 else
12 \hspace{1cm} result ← StaticFixed;
13 \hspace{1.5cm} foreach field $f$ of type $t_{arg}$ do
14 \hspace{2cm} tmp ← AnalyzeField($f$);
15 \hspace{2cm} if tmp = Variable then return Variable;
16 \hspace{2cm} else if tmp = RuntimeFixed then
17 \hspace{3cm} result ← RuntimeFixed;
18 end
19 return result;
20 end
21
22 Function AnalyzeField($f_{arg}$);
23 result ← StaticFixed;
24 \hspace{1cm} foreach runtime type $t$ in $f_{arg}$.getTypeSet do
25 \hspace{2cm} tmp ← AnalyzeType($t$);
26 \hspace{2cm} if tmp = Variable then return Variable;
27 \hspace{2cm} else if tmp = RuntimeFixed then
28 \hspace{3cm} if $f_{arg}$ is not final then return Variable;
29 \hspace{3cm} else result ← RuntimeFixed;
30 end
31 return result;
32 end
33 end
34

Two indirect-recursive functions, AnalyzeType (lines 4–22) and AnalyzeField (lines 23–34), are used to further determine the size-type of the target UDT. The stop condition of the recursion is when the current type is a primitive type (line 5). We treat each array type as having a length field and an element field. Since different instances of an array type can have different lengths, arrays with static fixed-sized elements will be considered as an RFST (lines 8–9).

We define a total ordering of the variability of the size-types (except the recursively-defined type) as follows: $SFST < RFST < VST$. Based on this order, the size-type of each UDT is determined by its field that has the highest variability (lines 12–20). Furthermore, each field’s final size-type is determined by the type with the highest variability in its type-set. But a non-final field of an RFST will be finally classified as VST, because the same field can possibly point to objects with different data-sizes (lines 28–29). Consider that whenever we find a VST field, the top-level UDT must also be classified as a VST. In this case, the function can immediately return without further traversing the graph.

We take the type LabeledPoint in Figure 1 as a running example. In Figure 3 every field has a type-set with a single element and the declared type of each filed is equal to its corresponding runtime type except that the features field has a declared type (Vector), while its runtime type is DenseVector. Moreover, for a more sophisticated implementation of logistic regression with high-dimensional data sets, the features field can have both DenseVector and SparseVector in its type-set.

Since there is no cycle in the type dependency graph, LabeledPoint is not a recursively-defined type. As shown in Figure 3 LabeledPoint contains a primitive field (i.e. label) and a field of the Vector type (i.e. features). Therefore, the size-type of LabeledPoint is determined by the size-type of features, i.e. the size-type of DenseVector. It contains four fields: one of the array type and other three of the primitive type. The data field will be classified as an RFST but not a VST due to its final modifier (val in Scala). Furthermore, the DenseVector objects assigned to features can have different data-size values because they may contain different arrays. Therefore, both features and LabeledPoint belong to VST.

3.3 Global Classification Analysis

The local classification algorithm is easy to implement and has negligible computational overhead. But it is conservative and often overestimates the variability of the target UDT. For example, the local classifier conservatively assumes that the features field of a LabeledPoint object may be assigned with DenseVector objects with different data-size values. Therefore it mistakenly classifies it as a VST, which can not be safely decomposed.

Furthermore, the local classifier assumes that the DenseVector objects contain arrays (features.data) with different lengths. Even if we change the modifier of features from var to val, i.e. only allowing it to be assigned once, the local classifier still considers it as an RFST rather than an SFST.

For UDTs that are categorized as RFST or VST, we further propose an algorithm to refine the classification results via global code analysis on the relevant methods of the UDTs. To break the assumptions of the local classifier, the global one uses code analysis to identify init-only fields and fixed-length array type according to the following definitions.

Init-only field. A field of a non-primitive type $T$ is init-only, if, for each object, this field will only be assigned once during the program execution. $^{1}$

$^{1}$We always treat the array element fields as non init-only, otherwise the analysis needs to trace the element index value
Fixed-length array types. An array type \( A \) contained in the type-set of field \( f \) is a fixed-length array type w.r.t. \( f \) if all the \( A \) objects assigned to \( f \) are constructed with identical length values within a well-defined scope, such as a single Spark job stage or a specific cached RDD. An example of symbolized constant propagation is shown in Figure 4. Here, \( \text{array} \) is constructed with the same length for whatever \( \text{foo()} \) returns. The fixed-length array types with its element fields being SFST (or RFST) can be refined to SFST (or RFST).

**Figure 4: Symbolized constant propagation.**

In Figure 4 the \text{features} field is only assigned in the constructor of LabeledPoint (lines 1–8), and the length of features.data is a global constant value \( \text{D} \) (lines 14–16). Thus, the size-class of LabeledPoint can be refined to SFST.

**Algorithm 2:** Global classification analysis.

**Algorithm 3:** Static fixed-sized type refinement: \( \text{SRefine}(t_{\text{arg}}, g_{\text{arg}}) \)

**Algorithm 4:** Runtime fixed-sized type refinement: \( \text{RRefine}(t_{\text{arg}}, g_{\text{arg}}) \)

from the outside of the call graph or returned by the I/O operations will be represented by symbols considered as constant values.

**3.4 Phased Refinement**

In a typical data parallel programming framework, such as Spark, each job can be divided into one or more execution phases, each consisting of three steps: (1) reading data from materialized (on-disk or in-memory) data collectors, such as cached RDD, (2) applying an UDF on each data object, and (3) emitting the resulting data into a new materialized data collector. Figure 5 shows the framework of a job in Spark. It consists one or more top-level computation loops, each reads data object from its source, and writes the results
into the sink. Every two successive loops are bridged by a data collector, such as an RDD or a shuffle buffer.

We observe that the data-sizes of object types may have different levels of variability at different phases. For example, in an early phase, data would be grouped together by their keys and their values would be concatenated into an array whose type is a VST at this phase. However, once the resulting objects are emitted to a data collector, e.g., a cached RDD, the subsequent phases might not reassign the array fields of these objects. Therefore, the array types can be considered as RSFTs in the subsequent phases. We exploit this phenomenon to refine a data type’s size-class in each particular phase of a job, which is called phased refinement. This can be achieved by running the global classification algorithm for the VSTs on each phase of the job.

4. LIFETIME-BASED MEMORY MANAGEMENT

4.1 The Spark Programming Framework

Spark provides a functional programming API, through which users can process Resilient Distributed Datasets (RDDs), the logical data collections partitioned across a cluster. An important feature is that RDDs can be explicitly cached in the memory to avoid re-computation or disk I/O overhead.

While Spark supports many operators, the ones most relevant for memory management are some key-based operators, including reduceByKey, groupByKey, join and sortByKey (analogues of GroupBy-Aggregation, GroupBy, Inner-Join and OrderBy in SQL). These operators process data in the form of Key-Value pairs. For example, reduceByKey and groupByKey are used for: 1) aggregating all Values with the same Key into a single Value; 2) building a complete Value list for each Key for further processing.

Furthermore, these operators are implemented using data shuffling. The shuffle buffer stores the combined value of each Key. For example, for the case of reduceByKey, it stores a partial aggregate value for each Key, and for the case of groupByKey, it stores a partial list of Value objects for each Key. When a new Key-Value pair is put into the shuffle buffer, eager combining is performed to merge the new Value with the combined value.

For each Spark application, a driver program negotiates with the cluster resource manager (e.g., YARN [34]), which launches executors (each with fixed amount of CPU and memory resource) on worker machines. An application can submit multiple jobs. Each job has several stages separated by data shuffles and each stage consists of a set of tasks that perform the same computation. Each executor occupies a JVM process and executes the allocated tasks concurrently in a number of threads.

4.2 Lifetimes of Data Containers in Spark

In Spark, all objects are allocated in the running executors’ JVM heaps, and their references are stored in three kinds of data containers described below. A key challenge for Deca is deciding when and how to reclaim the allocated space. In the lifetime analysis, we focus on the end points of the lifetime of the object references. The lifetime of an object ends once all its references are dead.

UDF variables. Each task creates function objects according to its task descriptor. UDF variables include objects assigned to the fields of the function objects and the local variables of their methods. The lifetimes of the function object end when the running tasks complete. In addition, as long-living objects are recommended to be stored in cached RDDs, in most applications, local variables are dead after each method invocation. Therefore, we treat all the data objects referenced only by the local variables as short-living temporal objects.

Cache blocks. In Spark, each RDD has an object that records its data source and the computation function. Only the cached RDDs will be materialized and retained in memory. A cached RDD consists of a number of cache blocks, each being an array of objects. The lifetimes of cached RDDs are explicitly determined by the invocations of cache() and unpersist() in the applications. Whenever a cached RDD has been “unpersisted”, all of its cache blocks will be released immediately. For non-cached RDDs, the objects only appear as local variables of the corresponding computation functions and hence are also short-living.

Shuffle buffers. A shuffle buffer is accessed by two successive phases in a job: one creates the shuffle buffer and puts data objects into it, while the other reads out the data for further processing. Once the second phase is completed, the shuffle buffer will be released.

With regard to the lifetimes of the object references stored in a shuffle buffer, there are three situations. 1) In a sort-based shuffle buffer, objects are stored in an in-place sorting buffer sorted by the Key. Once object references are put into the buffer, they will not be removed by the subsequent sorting operations. Therefore, their lifetimes end when the shuffle buffer is released. 2) In a hash-based shuffle buffer with a reduceByKey operator, the Key-Value pairs are stored in an open hash table with the Key object as the hash key. Each aggregate operation will create a new Value object while keeping the Key objects intact. Therefore, a Value object reference dies upon an aggregate operation over its corresponding Key. 3) In a hash-based shuffle buffer with a groupByKey operator, a hash table stores a set of Key objects and an array of Value objects for each Key. The combining function will only append Value objects to the corresponding array and will not remove any object reference. Hence, the references will die at the same time as the shuffle buffer. Note that these situations cover all the key-based operators in Spark. For example, aggregateByKey and join are similar.
to `reduceByKey` and `groupByKey` respectively. Other key-based operators are just extensions of the above basic operators and hence can be handled accordingly.

### 4.3 Data Containers in Deca

As discussed above, object references’ lifetimes can be bound with the lifetimes of their containers. Deca builds a data dependent graph for each job stage by points-to analysis [21] to produce the mapping relationships between all the objects and their containers. Objects are identified by either their creation statements if they are created in the current stage, or their source cached blocks if they are read from cached blocks created by the previous stage.

However, an object can be assigned to multiple data containers. For example, if objects are copies between two different cached RDDs, then they can be bound to the cached containers. In such cases, we assign a sole primary container as the owner of each data object. Other containers are treated as secondary containers. The object ownership is determined based on the following rules:

1. Cached RDDs and shuffle buffers have higher priority of data ownership than UDF variables, simply due to their longer expected lifetimes.

2. If there are objects assigned to multiple high-priority containers in the same job stage, the container created first in the stage execution will own these objects.

In the rest of this subsection, we present how data are organized within the primary and secondary containers under various situations.

#### 4.3.1 Memory Pages in Deca

Deca uses unified byte arrays with a common fixed size as logical memory pages to store the decomposed data objects. A page can be logically split into consecutive byte segments, one for each top-layer object. Each of such segment can be further split into multiple segments, one for each lower-layer object, and so on. The page size is chosen to ensure that there is only a moderate number of pages in each executor’s JVM heap so that the GC overhead is negligible. On the other hand, the page size should not be too large either, so that there would not be a significant unused space in the last page of a container.

For each data container, a group of pages are allocated to store the objects it owns. Deca uses a page-info structure to maintain the metadata of each page group. The page-info of each page graph contains: 1) `pages`, a page array storing the references of all the allocated pages of this page group; 2) `endOffset`, an integer storing the start offset of the unused part of the last page in this group; 3) `curPage` and `curOffset`, two integer values that store the progress of sequentially scanning, or appending to, this page group.

#### 4.3.2 Primary Container

The way how Deca stores objects in their primary container depends on the type of the container:

**UDF variables.** Deca does not decompose objects owned by UDF variables. These objects do not incur significant GC overheads, because: (1) the objects only referenced by local variables are short-living objects and they belong to the young generation, which will be reclaimed by the cheap minor GCs; (2) the objects referenced by the function object fields may be promoted to the part of old generation, but the total number of these objects in a task is relatively small in comparing to the big input dataset.

**Cache blocks.** Deca always decomposes the SFST or RFST objects and stores their raw data bytes in the page group of a cache block, while keeps the VST objects intact. Figure 6(a) shows the structure of a cache block of a cached RDD, which contains decomposed objects.

A task can read objects from a decomposed cache block created in a previous phase. If this task changes the data-sizes of these objects, Deca has to re-construct the objects and release the original page group. To avoid thrashing, when such re-construction happens, Deca will not re-decompose these objects again even if they can be safely decomposed in the subsequent phases.

**Shuffle buffers.** Figure 6(b) shows the structure of a shuffle buffer. Similar to cache blocks, data of an RFST or an SFST in a shuffle buffer will be decomposed into the shuffle buffer’s page group. However, unlike cached RDD, where data are accessed in a sequential manner, data in a shuffle buffer will be randomly accessed to perform sorting or hashing operations. Therefore, as illustrated on the left-hand side of Figure 6(b), we use an array to store the pointers to the keys and values within a page. The hashing and sorting operations are performed on the pointer arrays. However, the pointer array can be avoided for a hash-based shuffle buffer with both the `Key` and the `Value` being of primitive types or SFSTs. This is because we can deduce the offsets of the data within the page statically.

As we discussed in Section 4.2, for a hash-based shuffle buffer with a GroupBy-Aggregation computation, a combining operation would kill the old `Value` object and create a new one. Therefore, `Value` objects are not long-living and frequent GC of these objects are generally unavoidable. However, if the `Value` object is of an SFST, then we can still decompose it and whenever a new object is generated by the combining operation, we can just reuse the page segment occupied by the old object, because the old and the new objects are of the same size. Doing this would save the frequent GC caused by these temporary `Value` objects.

For brevity, we omit swapping data between memory and disks here. It is straightforward to adapt to the cases with disk swapping for data caching and shuffling (see Appendix C for details).

#### 4.3.3 Secondary Container

There are common patterns of multiple data containers sharing the same data objects in Spark programs, such as: 1) manipulating data objects in cache blocks or shuffle buffers through UDF variables; 2) copying objects between cached RDDs; 3) immediately caching the output objects of shuffling; 4) immediately shuffling the objects of a cached RDD.
We implement Deca based on Spark in roughly 6700 lines of Scala code. It consists of an optimizer used in the driver, and a memory manager used in every executor. The memory manager allocates and reclaims memory pages. It works together with the Spark cache manager and shuffle manager, which manage the un-decomposed data objects. The optimizer analyzes and transforms the code of each job when it is submitted in the driver (see Appendix A for details). The transformed code will use the API provided by the memory manager to create pages and access the stored bytes.

The Deca optimizer uses the Soot framework [6] to analyze and manipulate the Java bytecode. The Soot framework provides a rich set of utilities, which implements classical program analysis and optimization methods. The optimization consists of three phases: pre-processing, analysis and transformation.

In the pre-processing phase, Deca uses iterator fusion [27] to bundle the iterative and isolated invocations of UDFs into larger, hopefully optimized code regions to avoid complex and costly inter-procedural analysis. The per-stage call graphs and per-field type-sets are also built using Soot in this phase. Building per-phase call graphs will be delayed to the analysis phase if a phased refinement is necessary. In the analysis phase, Deca uses methods described in Section 3 and Section 4 to determine whether and how to decompose particular data objects in their containers. Based on the obtained decomposability information, new class files with transformed code will be generated and distributed to all executors in the transformation phase. In general, the accessing of primitive fields of decomposed data objects in the original code will be transformed to accessing the corresponding page segments (see Appendix B for details).

6. EVALUATION

We use five nodes in the experiments, with one node as the master and the rest as workers. Each node is equipped with two eight-core Xeon-2670 CPUs, 64GB memory and one SAS disk, running RedHat Enterprise Linux 5 (kernel 2.6.18) and JDK 1.7.0 (with default GC parameters). We use three real graphs: LiveJournal social network [13], Random Walk with Restart (PR), and Connected Component (CC). As shown in Table 1 they exhibit different characteristics and hence can verify the system’s performance in various different situations. For WC, we use the datasets produced by Hadoop RandomWriter with different unique key numbers (1M and 100M) and sizes (50GB, 100GB and 150GB). PR and KMeans use: 4096-dimension feature vectors (40GB and 80GB) extracted from Amazon image dataset [24] and randomly generated 10-dimension vectors (ranging from 40GB to 200GB). For PR and CC, we use three real graphs: LiveJournal social network [13] (2GB), webbase-2001 [14] (30GB) and a 60GB graph generated by HiBench [3]. The maximum JVM heap size of each executor is set to be 30GB for the applications with only data caching or data shuffling, and 20GB for those with both caching and shuffling.

6.1 Impact of Shuffling
WC is a two-stage MapReduce application with data shuffling between the “map” and “reduce” stages. We examine the lifetimes of data objects in the shuffle buffers with the smallest dataset. We periodically record the alive number of objects and the GC time with JProfiler 9.0. The result is shown in Figure 8(a). WC uses a hash-based shuffle buffer to perform eager aggregation, which is implemented in Tuple2. The number of Tuple2 objects, which fluctuates during the execution, can indicate the number of objects in shuffle buffers. While the number of Tuple2 is also large in “map” stage but decrease in shuffle in Deca. GCs are triggered frequently to release the space occupied by the temporary objects in the shuffle buffers.

To avoid such frequent GC operations, Deca reuses the space occupied by the partially-aggregated Value for each Key in the shuffle buffer. Figure 8(b) compare the execution times of Deca and Spark. In all cases, Deca can reduce the execution time by 10%-58%. One could also see that the performance improvement increases with more number of keys. This is because the size of a hash-based shuffle buffer with eager aggregation mainly depends on the number of keys. The reduction of GC overhead would become more prominent with a larger number of keys. Furthermore, since Deca stores the objects in the shuffle buffer as byte arrays, it also saves the cost of data (de-)serialization by directly outputting the raw bytes.

![Table 1: Applications used in the experiments](image)

<table>
<thead>
<tr>
<th>Application</th>
<th>Stages</th>
<th>Jobs</th>
<th>Cache</th>
<th>Shuffle</th>
</tr>
</thead>
<tbody>
<tr>
<td>WC</td>
<td>two</td>
<td>single</td>
<td>non</td>
<td>aggregated</td>
</tr>
<tr>
<td>LR</td>
<td>single</td>
<td>multiple</td>
<td>static</td>
<td>non</td>
</tr>
<tr>
<td>KMeans</td>
<td>two</td>
<td>multiple</td>
<td>static</td>
<td>aggregated</td>
</tr>
<tr>
<td>PR</td>
<td>multiple</td>
<td>multiple</td>
<td>static</td>
<td>grouped</td>
</tr>
<tr>
<td>CC</td>
<td>multiple</td>
<td>multiple</td>
<td>static</td>
<td>aggregated</td>
</tr>
</tbody>
</table>

![Figure 8: Results of shuffling-only WC exec](image)

(a) WC lifetime  
(b) WC exec

WC is a two-stage MapReduce application with data shuffling between the “map” and “reduce” stages. We examine the lifetimes of data objects in the shuffle buffers with the smallest dataset. We periodically record the alive number of objects and the GC time with JProfiler 9.0. The result is shown in Figure 8(a). WC uses a hash-based shuffle buffer to perform eager aggregation, which is implemented in Tuple2. The number of Tuple2 objects, which fluctuates during the execution, can indicate the number of objects in shuffle buffers. While the number of Tuple2 are also large in “map” stage but decrease in shuffle in Deca. GCs are triggered frequently to release the space occupied by the temporary objects in the shuffle buffers.

To avoid such frequent GC operations, Deca reuses the space occupied by the partially-aggregated Value for each Key in the shuffle buffer. Figure 8(b) compare the execution times of Deca and Spark. In all cases, Deca can reduce the execution time by 10%-58%. One could also see that the performance improvement increases with more number of keys. This is because the size of a hash-based shuffle buffer with eager aggregation mainly depends on the number of keys. The reduction of GC overhead would become more prominent with a larger number of keys. Furthermore, since Deca stores the objects in the shuffle buffer as byte arrays, it also saves the cost of data (de-)serialization by directly outputting the raw bytes.

### 6.2 Impact of Caching

LR and KMeans are representative machine learning applications that perform iterative computations. Both of them first load and cache the training dataset into memory, then iteratively update the model until the pre-defined convergence condition is met. In our experiments, we only run 30 iterations. We do not account for the time to load the training dataset, because the iterative computation dominates the execution time, especially consider that these applications can run up to hundreds of iterations in a production environment. We set 90% of the available memory to be used for data caching.

We first examine the lifetimes of data objects in cache RDDs for LogisticRegression (LR) using the 40GB dataset.

The result is shown in Figure 9(a). We find that the number of objects is rather stable throughout the execution in Spark, but full GCs have been triggered several times in vain (the peaks of the GC time curve). This is because most objects are long-living and hence their space cannot be reclaimed. While these objects are less in Deca because they are transformed to bytes after being read from the HDFS. Some objects still live in old generation of JVM heap because no full GC is active.

By grouping massive objects with the same lifetime into a few byte arrays, Deca can effectively eliminate the GC problem of repeatedly scanning alive data objects for their liveness. Figure 9(b) and Figure 9(c) show the execution times of LR and KMeans for both Deca and Spark. Here we also examine the cases using Kryo to serialize the cached data in Spark, which is denoted as “SparkSer” in the figures.

For the 40GB and 60GB datasets, the improvement is moderate and can be mainly attributed to the elimination of object creation and minor GCs. In these cases, the memory is sufficient to store the temporary objects, and hence full GC is rarely triggered. Furthermore, serializing the cached data also helps reducing the GC time. Therefore, with the 40GB dataset, SparkSer outperforms Spark by reducing the GC overhead. However, for larger datasets, the overhead of data (de-)serialization cannot pay off the reduced GC overhead. Therefore, simply serializing the cached data is not a robust solution.

For the three larger datasets the improvement is more significant. The speedups of Deca are ranging from 16x to 41.6x. In these datasets, the long-living objects consume almost all available memory space, and therefore full GCs are frequently triggered, which just repeatedly and unavailingy trace the cached data objects in the old generation of the JVM heap. With the 100GB and 200GB datasets, the additional disk I/O costs of cache swapping also prolong the execution times of Spark. Deca keeps a smaller memory footprint of cached data and swap smaller portion of data to the disks.

We also conduct the experiments on a real dataset, Amazon image dataset with 4096 dimensions. Figure 9(d) shows the speedups achieved by Deca are ranging from 1.2x to 5.3x. With such a high dimensional dataset, the size of object headers becomes negligible and therefore, the sizes of the cached data of Spark and Deca are nearly identical.

### 6.3 Impact of Mixed Shuffling and Caching

<table>
<thead>
<tr>
<th>Graph</th>
<th>LiveJournal (LJ)</th>
<th>WebBase (WB)</th>
<th>HiBench (HB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vertices</td>
<td>4.8M</td>
<td>118M</td>
<td>602M</td>
</tr>
<tr>
<td>Edges</td>
<td>68M</td>
<td>1B</td>
<td>2B</td>
</tr>
<tr>
<td>Data Size</td>
<td>26GB</td>
<td>30GB</td>
<td>60GB</td>
</tr>
</tbody>
</table>

![Table 2: Graph datasets used in PR and CC](image)
Figure 9: Results of caching-only LR/KMeans

Figure 10: Results of PR and CC

Table 3: GC time reduction.

<table>
<thead>
<tr>
<th>App</th>
<th>exec</th>
<th>gc</th>
<th>ratio</th>
<th>gc</th>
<th>reduction</th>
</tr>
</thead>
<tbody>
<tr>
<td>WC: 150GB</td>
<td>4895s</td>
<td>2016s</td>
<td>46.5%</td>
<td>12.2s</td>
<td>99.4%</td>
</tr>
<tr>
<td>LR: 80GB</td>
<td>2412s</td>
<td>2069.9s</td>
<td>73.4%</td>
<td>2.5s</td>
<td>99.9%</td>
</tr>
<tr>
<td>KMeans: 80GB</td>
<td>5443s</td>
<td>4299.8s</td>
<td>78.9%</td>
<td>7.2s</td>
<td>99.8%</td>
</tr>
<tr>
<td>PR: 30GB</td>
<td>5544s</td>
<td>3596.6s</td>
<td>64.7%</td>
<td>21.7s</td>
<td>99.4%</td>
</tr>
<tr>
<td>CC: 80GB</td>
<td>2088s</td>
<td>1443.9s</td>
<td>69.2%</td>
<td>36s</td>
<td>97.5%</td>
</tr>
</tbody>
</table>

6.4 GC improvement

Table 4 shows the times to run GC and the ratios of GC time to the whole job execution time for the seven applications. For each application, we only present the case with the largest input dataset that does not have data swapping or spilling, to avoid the disk I/O affecting the execution time. For each case, the GC time is an average of the values on all executors. The result demonstrates the effect of GC elimination by Deca, and how it improves the entire application performance.

As shown in the result, the GC running time of LR and KMeans occupies the largest portion of the total execution time among all cases, which are 73.4% and 78.9% respectively. With the 80GB input dataset, the cached data objects almost consume all the memory space of the old generation of the JVM heap. Deca reduces GC running time in two ways: 1) smaller cache datasets trigger much less full GCs; 2) once a full GC is triggered, the overhead of tracing objects is significantly reduced.

Since all the other applications have shuffle phases in their executions, the disk and network I/O account for a significant portion of the total execution time. Furthermore, reserving memory spaces for shuffle buffers makes that the long-living cached objects occupy no more than 60% of the total available memory. Therefore, in these cases the GC running time ranges from 64.7% to 69.2%. This explains the different improvement ratios for different types of applications reported above.

We then compare Deca with GC tuning methods. The Spark document [7] states that adjusting the fractions of memory allocated to cache blocks and to shuffle blocks is an effective GC tuning method. Furthermore, we also compare with two GC algorithms available in Hotspot JVM: namely CMS and G1. Table 4 shows the results. LR is very sensitive to GC tuning. By setting fractions of cache and shuffle buffer to 0.6 and 0.4 (the optimal based on our experiments), respectively, or replacing PS with CMS or G1 with tuned parameters, we can significantly improve the performance of LR. However, PR is much less sensitive to GC tunings, which is consistent with the previously reported experiments [6]. However, we cannot achieve the same performance gain by setting a higher number of concurrent GC threads in G1 as reported in [6]. We conjecture that it is because of the difference of the configuration of the machines, which is not stated in [6]. This experiment indicates that GC tuning is an effective way to improve GC performance in some applications, however it is a cumbersome process and is highly dependent on the applications and the system environment.

6.5 Microbenchmark

To make a closer comparison, we attempt to break down
programming framework.

In this experiment, we compare Deca with Spark SQL, which is optimized for SQL-like queries. Spark SQL uses a serialized column-oriented format to store in-memory tables, and, with the project Tungsten, the shuffled data of certain (built-in or user-defined) aggregation functions, such as AVG and SUM, are also stored in memory with a serialized form. The serialization can be either auto-generated or manually written by users. We use a dataset sampled from the Common Crawl document corpus, where rankings is 44GB. We use the table schemas and the two SQL queries provided in Spark’s benchmark [5]. For each query, a semantic-identical hand-written Spark program (with RDDs) are used for Spark and Deca. The input tables are entirely cached in memory before being queried. We disable the in-memory compression of Spark SQL.

The first query is a simple filtering:

```sql
SELECT pageURL, pagerank FROM rankings
WHERE pagerank > 100;
```

The second one is a typical GroupBy aggregate query:

```sql
SELECT SUBSTR(sourceIP, 1, 5), SUM(adRevenue)
FROM uservisits
GROUP BY SUBSTR(sourceIP, 1, 5);
```

The results are shown in Table 6. All three systems perform equally well for the simple filtering query with small input table. Although the GC running time in Spark is higher than that in the other two systems, it only accounts for a negligible portion of the total execution time. For the second query with a larger table, the GC overhead is significant for Spark. We can see that, similar to Spark SQL, Deca can reduce more than 50% of the execution time in comparing to Spark, while keeping the generality of Spark’s programming framework.
parallel systems make use of its specific computation structure to realize more complex memory management. Since the early adoption of JVM in implementing SQL-based data-intensive systems [30], efforts have been devoted to making use of the well-defined semantics of SQL query operators to improve the memory management performance in managed runtime platforms. Spark SQL [12] transforms relational tables to serialized bytes in a main-memory columnar storage. Tungsten [4], a Spark sub-project, enables the serialization of hash-based shuffle buffers for certain Spark SQL operators. Deca has a similar performance as Spark SQL for structured data processing, meanwhile it provides more flexible computation and data models, which eases the implementation of advanced iterative applications such as machine learning and graph mining algorithms.

8. CONCLUSION

In this paper, we identify that GC overhead in distributed data processing systems is unnecessarily high, especially with a large input dataset. By presenting Deca’s techniques of analyzing the variability of object sizes and safely decomposing objects in different containers, we show that it is possible to develop a general and efficient lifetime-based memory manager for distributed data processing systems to largely eliminate the high GC overhead. The experiment results show that Deca can significantly reduce Spark’s application running time for various cases without losing the generality of its programming framework. Finally, to take advantage of Deca’s optimization, a user is recommended to not creating a massive number of long-living objects of a VST, which cannot be safely decomposed.

9. ACKNOWLEDGMENTS

We want to thank Beng Chin Ooi of the NUS, Xipeng Shen of the NCSU, and Bingsheng He of the NTU for their valuable comments.

10. REFERENCES


Table 6: Execution times of two exploratory SQL query in Spark, Spark SQL and Deca. In Query2, the size of swapped cache data in Spark is 23.1GB.


APPENDIX

A. HYBRID OPTIMIZATION

Intuitively, Deca can be implemented as a standalone tool that transforms the compiled jar files of a Spark program before its execution. However, a Spark driver program may execute many jobs, each consisting of several stages separated by shuffles. The job submission will be implicitly triggered by an action, such as reduce, which returns a value to the driver after running a UDF on a dataset. According to the results returned by the current job, the driver decides how to submit the next job.

A driver program can freely use the control statements (if/for/while) to control the computation. Therefore, it may submit different jobs with different input datasets and configuration parameters. A static optimization has to enumerate and process all the possible jobs by exhaustively exploring a large number of possible execution paths of the program, which is the well-known path explosion problem. This is even infeasible if the program has loop structures, which renders the number of execution paths unbounded.

To address these challenges, we implement Deca in a hybrid way, which contains a static analyzer and a runtime optimizer. The static analyzer extracts priori knowledge about the UDFs and UDTs of the target programs, which can be used to reduce the runtime optimization overheads. The runtime optimizer intercepts the submitted jobs at runtime, and optimizes each job before actually submitting it to the Spark platform. With this approach, Deca only optimizes the actually submitted jobs, and thereby completely eliminates the need for exploring all the execution paths.

B. CODE TRANSFORMATION

In each stage, for the data objects of the UDT that can be safely decomposed, Deca transforms the corresponding code and leaves the unoptimizable part unchanged. The transformation phase can be further split into three sub-phases, which are described below.

Decomposition. In this sub-phase, Deca generates a synthesized class for each UDT (called SUDT) to access the decomposed raw data. Logically, the this reference of a decomposed UDT object will be transformed to the start offset (index of the first byte of its raw data) of its containing byte array. Every field accessing bytecode of this object will be transformed to the array accessing code based on the absolute field offset (object_start_offset + relative_field_offset).

The offset computation depends on the raw data size of each UDT instance. In each SUDT, Deca synthesizes static fields or methods to offer data-size values of all the UDT fields.

The data sizes of the primitive type fields are already defined in the official JVM specification, while the sizes of non-primitive type fields can be recursively got from the SUDTs of these fields. If a field data-size value can be determined, then it is stored as a global constant value in a static field in the corresponding SUDT. Otherwise, Deca synthesizes a static method of the SUDT to compute the data size during data processing. Similarly, for each UDT, Deca synthesizes static fields or methods to offer relative offset values of all
def computeGradient() = {
  val result = new Array[Double](D)
  var offset = 0
  while(offset < block.size) {
    var factor = 0.0
    val label = block.readDouble(offset)
    offset += 8
    for (i <- 0 to D) {
      var feature = block.readDouble(offset)
      offset += 8
      factor += weights(i) * feature
    }
    factor = (1 / (1 + exp(-label * factor)) - 1) * label
    offset -= 8 * D
    for (i <- 0 to D) {
      var value = block.readDouble(offset)
      offset += 8
      result(i) = result(i) + feature * factor
    }
  }
  result
}

Figure 12: The transformed code of Logistic Regression.

UDT fields in the SUDT. The relative offsets can be computed based on the field data sizes and the field order in the UDT definition. As an optimization technique, we can order the UDT fields by putting the fields with determinable sizes in the front. This method makes more field offsets values can be determined by Deca. Once all the field offsets can be computed, Deca transforms every methods of the UDT into its SUDT. During the transformation, all the field accessing code are replaced with the array accessing code.

Linking. Deca only transforms the code of UDTs in the decomposition sub-phase. In this sub-phase, Deca processes the code of UDFs and the main method for each stage. Deca memory manager allocates byte arrays rather than object arrays for data caching and shuffling. In Spark, every task processes data objects sequentially in each UDT object array in a loop. Because the array type is changed, the array index values in the loop must be computed based on the raw data sizes of each object in the optimized code. In fact, the array index variable stores the value of the start offset of the current processing data object. Based on the new array index values, Deca changes the code of UDFs and the main method in the following ways: 1) removes the invocations of UDT object constructors and directly write the initial values in the constructor parameters into the byte arrays based on the absolute field offset; 2) replaces all the field accessing code with the array accessing code; 3) replaces each invocation of a UDT method with the corresponding SUDT method, and adds the byte array and the object start offset as the additional parameters of the invocation.

Optimization. After decomposition and linking, the Deca optimizer further performs classical program optimizations, such as constant/copy propagation, method inlining and loop fusion. Most JVM JIT compilers prefer smaller methods, but we use method inlining aggressively in the pre-processing and transformation phases. If we produce giant methods that have sizes exceeding a threshold value, we split each giant method into several smaller ones using program slicing techniques.

While the code produced by Deca are JVM bytecode, in Figure 12 we provide the equivalent Scala code to illustrate the logic of the transformed LR code. The illustrated part is corresponding to the gradient computation of the original code in line 21-24 of Figure 1.

C. DISK SWAPPING/SPILLING

When the working set size is larger than the available memory space of an executor, Spark moves part of its data out of the memory. For cached RDDs, Spark uses the LRU strategy to select the cache blocks for eviction. The evicted data will be directly discarded or swapped to the local disk according to the user-specified storage-level. For shuffles, Spark always spills the partial data into temporary files, and merges them into final files at the end of task executions.

For cached RDDs, Deca modifies the original LRU strategy to evict page groups rather than cache blocks. Accessing in-page data through either page-infos or pointers will refresh the corresponding page group’s recently-used counter. Spark serializes cache block data before write them into disk files, or transfer them through network for non-local accesses. In Deca, the decomposed data bytes can be directly used for disk and network I/O.

For shuffles, like Spark, Deca sorts the pointers before spilling, and writes the spilled data into files according to the order of the pointers. If a shuffle buffer has only pointers that reference page segments, Deca does not spill these pointers because normally they only occupy a small memory space. It pauses the shuffling and triggers the cache block eviction to make enough room. Deca uses a small memory space (normally only one page) to merge sorted spilled files. Once the merging space is fully filled, the merged data will be flushed to the final output file.