FAKULTÄT FÜR INFORMATIK

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Efficient Data Mining Algorithms in Main Memory Databases
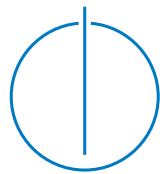
Martin Kapfhammer

FAKULTÄT FÜR INFORMATIK

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Efficient Data Mining Algorithms in Main Memory Databases

Effiziente Data Mining Algorithmen in Hauptspeicher-Datenbanken

Author:	Martin Kapfhammer
Supervisor:	Prof. Alfons Kemper, Ph.D.
Advisor:	Linnea Passing, M.Sc.
Submission Date:	15.04.2015

I assure the single handed composition of this master's thesis in informatics only supported by declared resources.

Munich, 15.04.2015   Martin Kapfhammer

Acknowledgments

Abstract

Contents

1. Introduction

Databases are more in the center of innovation than ever. With the growing demands on storing and analyzing large amounts of data, linked with the capabilities of modern hardware, database vendors and researchers face new challenges and possibilities. One of these challenges is to execute data mining algorithms directly on the database. In this thesis, an approach is presented to integrate data mining techniques into the main memory database HyPer.

1.1. Motivation

Traditionally, databases are disk-based systems separated into two parts: One system used for Online Transactional Processing (OLTP), optimized for high rates of mission-critical, transactional write requests. As second system, a data warehouse is used for Online Analytical Processing (OLAP), executing long-running queries to gain insight into the collected data, used to make future decisions upon. Due to this contradicting requirements of critical, short write requests on the one side and long running, business-intelligence gaining read requests on the other side, traditionally two separated systems are used. The data synchronization between the two systems is ensured by an Extract-Transform-Load (ETL) process: Data is extracted from the OLTP system, transformed and loaded into the OLAP system. Since this process implicates heavy load on the database, it is usually done periodically, e.g. once a night. Obviously, this architecture reveals several drawbacks, like stale data, redundancy and the maintenance cost of two systems.

Modern hardware allows us to move away from this paradigm: Instead of using disk-based databases the data can be stored in memory. Since memory can be accessed much faster than the disk, an unprecedented throughput of OLTP transactions is possible. Using snapshots of the transactional data by exploiting the virtual memory management of the operating system, OLAP queries can run in parallel next to the OLTP transactions on up-to-date data. Such a system is HyPer [KN11], a relational main memory database guaranteeing the ACID properties, actively developed at the Chair of Database Systems at the TU München, and the main system used for our research.

The possibility of executing OLAP queries on a relational DBMS without interfering

the OLTP transaction throughput opens new possibilities for database systems: Additionally to long-running OLAP queries, data mining algorithms can be implemented and executed in HyPer. Data mining extends the possibilities of OLAP by applying more complex algorithms to discover interesting patterns and knowledge from the data. Crucial techniques are mining frequent patterns and association rules, classification, clustering and outlier detection.

Usually data mining tools have to import and integrate data from databases and other data sources by ETL processes, resulting in the same disadvantages as for data warehouses: data is not up-to-date for the crucial analyzing phase and redundant. By implementing data mining algorithms right into the database, data analysts, data scientists and business executives gain huge benefits. Instead of bringing the data to the algorithm, the algorithms are brought to the data. No ETL processes are necessary to export data to an additional system for data mining. Instead, the algorithms can be executed directly on the database system on up-to-date data. The execution of the algorithms benefits from the modern hardware and high computational power of the database system. Additionaly, the data mining process can be combined with already existing database operators such as grouping and aggregation, resulting in a very natural environment for data analysis.

1.2. Research Questions

With more and more data generated by modern IT systems, the needs for analyzing large amounts of data are rapidly growing and data mining software gets an increasing amount of attention. So far, most data scientists use standalone data mining tools such as R [R D08], Julia [Bez+12] or the Python ecosystem SciPy [J+01]. Furthermore, the Java frameworks Weka [Hal+09] and ELKI [AKZ08] are used actively, in particular in the research community. All tools are providing an environment for statistical computing and the application of various data mining algorithms. Additionaly, for mining tera- and petabytes of data, scalable software such as Apache Hadoop [Shv+10a], Apache Spark [Zah+10] and the data mining and machine learning framework Apache Mahout [Apa] are used.

While all of the presented environments are frequently used by data scientists, they demonstrate one decisive drawback: Before executing data mining algorithms, the data first has to be fetched from the database and transformed into a format readable by these tools. Therefore a first enhancement is to bring the algorithms closer to the database. Tools like MADlib [Hel+12] and RapidMiner [Rap14] are supporting specific database engines and trying to fill this gap. Algorithms can be run on the database, using extensible SQL, already existing SQL operators, stored procedures and

the possibility to connect those with high-level "glue" code. These tools provide a middleware between the database and data mining software and therefore data export becomes obsolete.

However, for complex algorithms SQL operators have to be combined with high-level constructs, often resulting in bad performance. Therefore, database vendors such as Oracle and SAP are going one step further and providing their databases with more and more statistical and data mining functionalities using the existing SQL syntax and graphical user interfaces [Ora15] [SAP14]. This leads to several advantages: A database system provides already efficient data storage and access, therefore data mining algorithms implemented on the database can benefit from these data structures. Besides, databases are optimized for modern hardware, e.g. multi core processors and cache hierarchies, which makes them presumably faster than platform-independent tools. Furthermore, data mining algorithms can profit from database features such as parallelization, scalability, recovery and backup facilities as well as from the query language SQL. SQL itself comes already with useful algorithms for data analysis such as selection, sorting and aggregation. Therefore an extension of the query language to integrate other algorithms for data mining would feel very natural to the data scientist. Regarding those advantages our research goal is to extend HyPer by data mining functionalities that can be executed directly on the database, exploiting the performance of modern database hardware for computational operations and building a general-purpose system for OLTP, OLAP and data mining queries. This thesis presents a proof-of-concept implementation of data mining in HyPer, and is part of a larger project that goes one step further and will provide an intuitive user interface around the algorithms: A functional language that is easy to learn, avoids side-effects and provides plotting options.

1.3. Approach

As proof-of-concept, the well-known k-Means algorithm will be implemented as a HyPer operator. K-Means is one of the most popular clustering algorithms and available on all presented platforms. As the algorithm is relatively simple and straightforward, a good comparability between different tools and their performance is given. Since most data mining algorithms are non-deterministic, the time per iteration will be used as the main performance criterion. An implementation of k-Means will demonstrate the possibilities of implementing and running data mining algorithms on HyPer.

HyPer provides a programming model for the implementation of operators by combining pre-compiled C++ code with dynamically generated LLVM code. Yet, the programmer has a lot of freedom in implementing complex data mining operators.

This work compares different implementation aspects of the k-Means algorithm and aims to detect best practices, patterns and building blocks for the implementation of further algorithms.

The remainder of this thesis is organized as follows: In the next section, fundamentals of the HyPer database are presented. In Chapter 3, a brief introduction of the terms knowledge discovery and data mining is given. Then, the k-Means clustering algorithm is introduced. Chapter 5 discusses other data mining tools and databases with data mining functionalities. In Chapter 6 the implementation of the k-Means clustering algorithm as HyPer operator is depicted. Chapter 7 discusses extensive experiments that demonstrate the applicability of data mining algorithms on HyPer. Chapter 8 concludes the thesis.

2. The HyPer Main Memory Database

In this chapter, the relational main memory database system HyPer [KN11] is introduced. HyPer is a database system combining both OLTP and OLAP processing and is the main research subject of this work: Extending HyPer in order to execute data mining algorithms directly in the relational database.

2.1. Motivation

As already mentioned, historically databases are disk-based systems separated into two parts: An Online Transactional Processing (OLTP) system, optimized for high rates of mission-critical, transactional write requests, and an Online Analytical Processing (OLAP) system, executing long-running queries. The data synchronization is ensured by an ETL process that extracts, transforms and loads data from the OLTP to the OLAP system, which is usually done periodically.

Obviously, this architecture reveals several drawbacks. The periodical execution of the ETL process results in stale data on the OLAP system, e.g. when implementing a nightly ETL process, data can be outdated for up to 24 hours which is problematic for real-time data analysis. Furthermore, two systems lead to higher costs: Hardware and software costs must be taken into account as well as maintenance and incident management. Additionally, implementing an ETL process can make the two systems overly complex in contrast of having one system.

Addressing these challenges the Hybrid OLTP & OLAP High-Performance Database System (HyPer) was developed, with the goal to process OLTP transactions with high performance and throughput, and, on the same system, process OLAP queries on up-to-date data.

2.2. The HyPer Architecture

In this section we give an overview about the HyPer system architecture and important design decisions. The argumentation in this section is based on [KN11].

2.2.1. High-Throughput OLTP Processing

HyPer achieves major performance gains by omitting typical disk-based database characteristics that are not necessary anymore when data resides in main memory. For example, database-specific buffer management and page structuring is not needed. Instead, data is stored in main memory optimized data structures within the virtual memory. Hence, HyPer can use the highly efficient address-translation of the operating system without any additional indirection.

Historically, OLTP transactions are suffering from slow disk access and I/O time caused by disk-based database systems. Therefore, OLTP transactions are executed in parallel to exploit the waiting time and to enhance the performance of the system. In contrast, OLTP processing on HyPer is very fast because all the data is already loaded into main memory and slow disk access is omitted and not a bottleneck anymore. Therefore, transactions do not have to wait for I/O and HyPer can implement OLTP transactions as a single-threaded approach and executes all transactions sequentially. This is possible because the execution time of an OLTP transaction is in microseconds and even without parallel execution of transactions, a high-throughput is achievable. Experiments on the TPC-C benchmark, a OLTP performance benchmark for relational databases, show that HyPer is cabable of 100,000 transactions per second [Müh+14]. Another advantage of this simple model is that locking and latching of data objects becomes redundant since only one transaction is active for the entire database.

Even though HyPer achieves a high-throughput with sequential execution, these constraints can be relaxed when the database tables are partitioned. An example is multi-tenancy [Aul+08][Aul+09]: Several clients are using the database, but do not have access to the data of the other clients. Then, the OLTP transactions can be executed in multiple threads to exploit the multi core database hardware. Still, no locking and latching is necessary since all transactions are executed sequentially within its partition.

2.2.2. Long-Running OLAP Queries

Obviously, this sequential execution of transactions is only possible if there are not any long-running queries. Long-running queries would be a bottleneck for the entire database and all following queries have to wait until its termination. Therefore this section discusses how HyPer implements OLAP queries on up-to-date data without negatively affecting the OLTP transaction performance.

HyPer considers OLAP queries as a new process and creates a snapshot of the virtual memory of the OLTP process for its OLAP processes. In Unix-based systems, this is done by `forking` the OLTP process and creating a `child` process for the OLAP process, exploiting the virtual memory functionality of the operating system. The OLAP

process takes an exact copy of the OLTP system on process creation and is now able to execute long running queries without interfering the OLTP transaction throughput. Thanks to modern operating systems the creation of snapshots is very efficient because the memory segments are not physically copied. Instead, operating systems apply a `copy-on-update` strategy. That means that both OLTP and OLAP processes are sharing the same physical main memory location since their virtual address translation maps to the same segments. Therefore copying memory on creation is not necessary. Only if an object gets updated by the OLTP process, a new page has to be created for the OLTP process, while the OLAP virtual memory page is still the one available when the process was created. Since this mechanism is supported by hardware, it is very fast and efficient without any implementation overhead for the HyPer database system. In face, experiments have shown that a `fork` can be achieved in several milliseconds, which is almost independent of the size of the OLTP system.

Since OLAP queries are read-only HyPer can execute them in parallel in multiple threads sharing the same snapshot. As in the sequential execution, locking and latching is not necessary because the used data structures are immutable. This inter-query parallelization can tremendously speed up query processing on multi core computers. Another approach is the creation of multiple OLAP session by forking the OLTP memory periodically. Therefore, for each OLAP session a new virtual memory snapshot is created and used as the current snapshot for the new session. This allows parallelization not only as inter-query parallelization but also among the different OLAP sessions.

In this chapter we have shown the advantages of a modern main memory system such as HyPer. Without the bottleneck of disk I/O, database operations can be run sequentially with up to 100,000 TPC-C transactions per second. OLAP queries can run very efficiently in parallel to the OLTP transaction system on different virtual memory snapshots with access to up-to-date data.

With these promising results for OLTP and OLAP processing, the next step is to extend HyPer by data mining functionalities. Then, HyPer provides even more possibilities for efficient, real-time data analysis on a running transactional database. In Chapter 6 we discuss the implementation of data mining algorithms, including how HyPer implements operators in general using the `consume produce` programming model and LLVM code generation.

3. Knowledge Discovery and Data Mining

In this chapter we clarify the terms and the importance of knowledge discovery and data mining. In a nutshell it is the process of discovering knowledge and patterns from data. Therefore, many tools and algorithms exist which we describe in the following sections.

3.1. Motivation

As already mentioned, data growth is exploding and new data is generated every day by every aspect of daily life, such as businesses, society, science and medicine. With this huge amount of data available, it has become one of the most valuable resources of our decade, and in the media, data is often described with the importance of oil[1] or as the new currency of our decade[2].

However, before data has any value it has to be transformed into knowledge. For example, e-commerce companies are very keen to find out not only what their customers have bought but also what they are likely to buy in the future. This knowledge is then used to generate more revenue. A very well known example is amazon.com which is advertising a variety of similar products the customer recently looked for, or products another customer with similar characteristics has purchased in the past. Therefore data mining techniques are used to boost consumerism. Additionally, this knowledge might be used to optimize storage cost, e.g. by knowing how much of certain products will be needed to distribute in the next weeks.

Other areas for data mining are telecommunication network carriers with a huge amount of data traffic generated every day, the health and medical industry and the Web in general, to name just a few examples. Again it is all about knowledge. The telecommunication network carriers are interested in customer usage behavior, while state authorities are interested in wiretapping of potential criminals. Medical data might be used to install an ameliorated patient monitoring and Web data can be used for search engines to find the best results or for ad networks to deliver promising ads. The list of sources to generate data from and the knowledge one is interested in is

---

[1]http://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/, 15.03.2015.
[2]http://www.europeanvoice.com/event/data-the-new-currency/, 15.03.2015.

endless.

3.2. Knowledge Discovery

After understanding the importance of data mining for gaining knowledge and using this knowledge for further decision making, this section defines data mining in greater detail. The term data mining itself often leads to confusion - we are not mining data but instead we are looking for knowledge and interesting patterns in a given data set or database. Therefore, the term knowledge discovery is often more appropriate.

However, the usage of knowledge discovery and data mining as synonyms is not entirely accurate. Instead, data mining is often seen as a part of the knowledge discovery process as depicted in Figure 3.1. This process starts with the preprocessing steps data cleaning, integration and reduction. Afterwards, data mining and its algorithms are applied and the discovered patterns are evaluated and presented.

In this work we use the term data mining as described in Figure 3.1 since we are mainly interested in the algorithms applied in the data mining phase. Nevertheless, future work will cover the entire knowledge discovery chain and implement techniques to clean the data as well to present the results. Therefore, this section gives a profound overview over all the knowledge discovery steps; a focus on the data mining algorithms follows in Section 3.3.
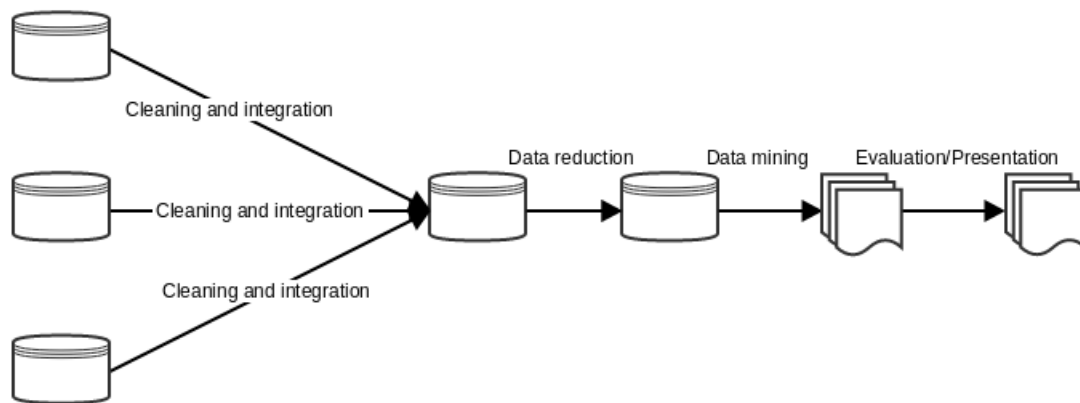
Figure 3.1.: The Process of Knowledge Discovery.

Regarding all those steps, Han et al. [HKP11, p. 8] give a detailed definition of data mining:

"Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically."

In the following sections we discuss each phase in detail and start with data cleaning.

3.2.1. Data Cleaning

The first step of knowledge discovery needs to ensure that the data we are interested in is complete, clean and consistent. Real world data is usually none of it. Data sets are incomplete, i.e. values are missing. If these missing values are crucial for the further data mining process, several data cleaning techniques exist to fill the gaps, e.g. by using an average value such as the mean or the median.
An example of inconsistent data are distinct values with the same meaning. Often, domain knowledge is required to transform the inconsistency into consistent data. Additionally, we often see random error or variance in a measured variable. This is called noisy data and must be cleaned for proper data analysis e.g. by smoothing the data set, using binning techniques.

3.2.2. Data Integration

Data mining often requires data from several sources. These sources can be databases, data warehouses, transactional data or plain data files. Also complex data might be used, such as time-related data, data streams, spatial data, multimedia data and graph data. All of these sources have to be integrated into one coherent data set which is then used for further analysis. There are several challenges regarding data integration, the most important is the entity identification problem: When combining data from different sources same data objects can have different names and types in different schemas. Therefore, data integration often requires domain knowledge of the different data sources and techniques used in the data cleaning phase to get a clean, consistent, non-redundant data set.

3.2.3. Data Reduction

After cleaning and integrating data into a coherent data set, the data is ready for further mining. However, those data sets are often of huge size and contain much information, not all necessary for the data mining process. Data reduction techniques can help to reduce the information density and speed up the data analysis and mining process. Simple techniques for data reduction are projection, selection and aggregation, available

in every database system. More advanced techniques are dimensionality reduction, e.g. by a principal component analysis, numerosity reduction above aggregation, e.g. clustering or sampling, and data compression.

Other methods are data transformations which change the data by converting values into another format e.g. by using normalization. Another possibility is data discretization, e.g. by transforming numeric data into intervals.

The aim of all of the addressed methods is to reduce the size of the initial data set and therefore to make it easier to get the information the user is interested in. It's also worth to mention that some of the presented data reduction techniques are data mining techniques as well, e.g. clustering. Therefore the boundary between data reduction and data mining is not always easy to set.

3.2.4. Data Mining

After applying the presented preprocessing steps the data set is now ready for data mining techniques. In the data mining phase we apply algorithms for finding interesting patterns in the data set and use techniques such as frequent pattern and association rule mining, classification, clustering and outlier detection. The algorithms for performing those techniques are presented in Section 3.3.

3.2.5. Pattern Evaluation

After generating those patterns we have to evaluate them to figure out how valuable the acquired knowledge is. There exist several criteria for pattern evaluation. First, a pattern has to be understandable. A clustering for example does not give any knowledge if the reason for the clustering is not comprehensible. That leads to the next criteria of evaluation: A pattern has to be valid, useful and novel. Only then, a pattern can be called interesting and can be used as knowledge, e.g. by fulfilling a hypothesis the user tries to confirm.

3.2.6. Knowledge Presentation

Knowledge presentation is the final step of knowledge discovery. Interesting patterns have to be described and visualized to the user in an appealing, understandable format. Apart from textual explanation, several data visualization techniques exist and should be applied for the respective situation.

3.3. Algorithms of Data Mining

In the following sections we discuss data mining techniques such as frequent itemset mining, classification, clustering and outlier detection. For each group the purpose of the technique is given and its most important algorithms are presented.

3.3.1. Mining Frequent Patterns and Association Rules

Frequent pattern mining aims to find the most frequent items in a data set and its association rules. A typical example is the market basket analysis: Supermarkets and e-commerce stores are interested in what their customers are buying, in particular in items that are commonly bought together. In other words: Which items are frequently put in the same market basket?
This knowledge is very valuable: It is not only about finding out which items are very popular and therefore get a better a position in the market or website. It is also about knowing which items are bought together which can lead to interesting decisions: Related items could be placed next to each other to make the shopping experience for the customer more convenient. Another approach could be to put related items far away from each other, so that the customer has to walk around in the store and is more likely to buy an additional product. These are just two possible strategies that could be applied with the knowledge resulting from frequent pattern mining.
Most algorithms for frequent pattern mining expect transactional data as input, because this type of data represents a market basket the best. The most popular algorithm is the Apriori algorithm [AS94], working in an iterative manner. Apriori generates itemsets of length k and checks if these itemsets appear frequently, i.e. if their count exceeds a given threshold. All the valid itemsets are then used in the next iteration to generate itemsets of length $k + 1$. The algorithm converges if no itemsets can be generated for the respective length.
Since the candidate generation and the scanning of the database is expensive, there are attempts to improve the Apriori performance [SON95][PCY95], as well as to establish other techniques. One of them is Frequent-Pattern-growth (FP-growth) [Han+04], an algorithm compressing the database into an FP-tree and therefore finding frequent itemsets without itemset generation.

3.3.2. Classification

In data mining, classification uses existing data as knowledge for predicting future events. A typical example is the identification of spam emails. This is done by labeling existing emails and categorizing them as spam or non-spam. Such a data set is called

the training data set. The algorithm uses then this training knowledge to classify new, unknown emails as spam or non-spam.

Since classification needs training data it is also called *supervised learning*. That means that before the start of the data mining process, previous knowledge has to be acquired. In our example, someone has to label existing emails first as spam or non-spam before the algorithm can be applied to unknown, unlabeled emails.

There are many classification algorithms in practice, popular ones are Support Vector Machines (SVM) [BGV92], Decision Trees [Qui93] and Neural Networks [Ros58]. An SVM classifier starts by building a model upon a training data set. Each data tuple is represented as a vector in the SVM model space. Since the data is labeled the SVM knows which data tuples belong together and tries to find separating hyperplanes between them. These hyperplanes can be described by a small subset of the data vectors called the *support vectors*. This fact makes the SVM very efficient on high-dimensional data. Unknown data tuples are then modeled as a vector in the SVM space and belong to the same category as the other data tuples they are separated together with by the hyperplanes.

3.3.3. Clustering

While classification requires apriori knowledge, i.e. tuples must be labeled to be used as training data, clustering does not require any previous knowledge. In that sense, clustering is a very convenient method to gain knowledge about a data set without the need of knowing exactly what the result should look like. Therefore, clustering is also called *unsupervised learning*.

An example is Google News, presenting and grouping news headlines obtained from many news websites. There is no predefined set of available news topics, instead the topics can change daily. Therefore, using classification with training data is not feasible. Instead, clustering can be used to find similar topics within the latest news articles and group them together. The clustering algorithm tries to find the best grouping of news, meaning that the news in one group have a high similarity with each other, while they are very different to the news in other groups.

Clustering techniques provide a variety of algorithms. These algorithms can be categorized in the following four groups:

- Partitioning methods: Partitioning methods are the most popular clustering algorithms, trying to find clusters of spherical shape. A distance function is used to measure similarity and dissimilarity among the clusters. Popular algorithms are k-Means [Llo82] and k-Medoids [KR90].

- Hierarchical methods: Hierarchical methods are useful for data consisting of sev-

eral hierarchies or levels. Clustering is applied by going up or down the hierarchy tree by merging or splitting subtrees, respectively. Similar to partitioning-based methods, hierarchical clustering algorithms tend to find spherical clusters.

- Density-based methods: For finding clusters of arbitrary shape, such as oval or "*S-shaped*" clusters, partitioning and hierarchical techniques are limited. Density-based clustering algorithms obtain much better results, using dense regions in data sets for identifying clusters instead of distances from a center point. The most popular algorithm is the DBSCAN (Density-based spatial clustering of Applications with Noise) [Est+96] algorithm. It works by finding core objects, i.e. objects within a dense neighborhood and joining those core objects together to find dense regions, i.e. clusters.

- Grid-based methods: All the presented methods so far are data-driven, i.e. groupings are detected by the distribution of objects in the embedding space. In contrast, grid-based methods are space-driven, i.e. the object space is mapped to a finite number of cells, resulting in very fast processing times for multi-dimensional data. Popular algorithms are STING [WYM97] and CLIQUE [Agr+98]. CLIQUE is a special form of a grid-based algorithm, since it is also density-based.

3.3.4. Outlier Detection

Outlier detection techniques aim to find data objects that behave in a different way than the majority of data objects. For an e-commerce system outliers can be clients spending much more money than the average client. This information is very valuable since the company is eager to make this client particularly satisfied, e.g. with special customer care treatment. Also in fraud detection and medical systems, outlier detection is very important to observe a security breach or a patient problem as early as possible.
Obviously there is a strong correlation between outlier detection and clustering. Data objects that do not fit into a cluster are potential outliers. Therefore clustering techniques can be applied for outlier detection. However, their main purpose is to find clusters, i.e. dense areas instead of outliers. Often we have to optimize these algorithms to omit dense areas to find outliers in a more efficient way.
Apart from *unsupervised learning*, *supervised learning* can be applied for outlier detection as well. First, data is labeled as normal or outlier, and upon that information future data can be classified. This data is then used as training data to classify data sets as outliers or normal data. The challenge of using classification methods for outlier detection is that outliers are very rare by definition, thus finding training instances resembling outliers is not trivial. Therefore typical classification algorithms often have to be adapted and optimized for outlier classification as well.

Apart from clustering and classification techniques, statistical and proximity-based methods are also used for outlier detection.

The remainder of this work discusses and implements the clustering algorithm k-Means, one of the most well-known and popular data mining algorithms. It is implemented in most modern data mining tools such as R and Julia, therefore a good comparison between different tools can be achieved. The algorithm is introduced in Chapter 4 and implemented in Chapter 6. The results are then discussed in Chapter 7.

4. The k-Means Clustering Algorithm

In this chapter the data mining algorithm k-Means is presented. K-means is a well-studied clustering algorithm, partitioning a data set into k clusters, where all data objects within a cluster are similar to each other and dissimilar to the data objects in the other clusters. The goal is to find the best clustering, minimizing the total distance between each data point and its assigned center. While the solution to that problem is NP-hard [MNV09][Alo+09], there are several heuristics, in particular the Lloyd algorithm [Llo82], which is a local search solution to this problem.

4.1. Motivation

We decided to use k-Means as proof-of-concept algorithm for data mining in the HyPer database because it is widely used and very popular. Furthermore, it is an iterative algorithm, which is very common for data mining algorithms. In fact, a survey of clustering data mining techniques in 2002 states that the algorithm "is by far the most popular clustering algorithm used in scientific and industrial applications" [Ber02]. K-Means is also part of the 10 top data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) [Wu+08], next to other famous algorithms such as the SVM [BGV92] and the PageRank [Pag+99] algorithm.
Since the algorithm is popular and easy to understand, k-Means is suitable as first implementation of a data mining algorithm on HyPer. To our best knowledge, all major data mining tools are implementing k-Means, which is a basic requirement regarding the comparability among existing tools and our own implementation. Hence, the running time is a good measurement for comparing the performance; the cluster compactness, i.e. the sum of squared errors, can be used as general criterion for the formal correctness of the implementation. Parts of k-Means can be executed in parallel, and since HyPer supports parallel computation, it will be interesting to evaluate how a single-threaded execution of k-Means performs against a multi-threaded computation. In fact, we are expecting tremendous performance gains executing k-Means in parallel on a multi core machine.

4.2. The Lloyd Algorithm

In this section we discuss the Lloyd algorithm [Llo82], usually referred as k-Means in literature and in this work as well. Formally, the k-Means problem and the k-Means algorithm are described the following: For a given integer k and a data set of n data points $X \subset \mathbb{R}^d$, choose k centers C to minimize the sum of squared error function,

$$sse = \Sigma_{x \in X} min_{c \in C} ||x - c||^2.$$

After finding these center points, each data point is assigned to its closest center which forms the clustering.

Solving such a problem is NP-hard, however Lloyd proposes a local search solution usually resulting in good groupings. The algorithm starts with k arbitrary center points, typically chosen at random from the data points. Then, each data point is assigned to the closest center using the euclidean distance. After assigning each data point, the centers get updated, i.e. the new center is the mean coordinate of all the data points assigned to this center. Then the next iteration begins, assigning the data points again, now to the updated centers. This continues until the process stabilizes and the algorithm converges.

Formally, the k-Means algorithm is described as the following:

1. Arbitrarily choose k centers $C = \{c_1, c_2, \cdots, c_k\}$ uniformly at random from X.

2. For each $i \in \{1, \cdots, k\}$, set the cluster C_i to be the set of points in X that are closer to cluster c_i than to any other cluster c_j for all $j \neq i$.

3. For each $i \in \{1, \cdots, k\}$, set c_i as the center of all points assigned to C_i:

$$c_i = \frac{1}{|C_i|} \Sigma_{x \in C_i} x.$$

4. Repeat Steps 2 and 3 until C no longer changes, i.e. the algorithm converges.

This simple algorithm terminates in practice very fast and provides mostly good results. However, the clustering can be arbitrarily bad for specific data sets. Therefore, MacQueen [Mac67] and Hartigan and Wong [HW79] propose improvements over the Lloyd algorithm, often leading to a faster convergence. Since the goal of this work is to demonstrate k-Means as a proof-of-concept implementation on HyPer, and since most tools, such as R and Julia, are implementing the Lloyd algorithm, we decided to use the Lloyd algorithm in the remainder of this work.

4.3. An Example of the k-Means Algorithm

In this section the k-Means algorithm is presented by a made-up example for deeper understanding. Our example data set consists of five data points with two dimensions: $x_1(3,9), x_2(2,5), x_3(5,8), x_4(7,5), x_5(4,2)$. This data will be clustered using the k-Means algorithm with $k = 2$, i.e. we are searching for two clusters in the data set. First, the algorithm starts with an initialization phase: Randomly, we choose two instances as initial center points. In this example, these data points are $c_1(3,9)$ and $c_2(4,2)$.

Next, we begin the first iteration of the k-Means algorithm and compute the euclidean distance from each data point to c_1 and to c_2. Each data point is then assigned to the closest cluster, as shown in Table 4.1. After Iteration 1, the data points x_1 and x_3 are closer to c_1 and therefore assigned to Cluster C_1, while the data points x_2, x_4 and x_5 belong to Cluster C_2.

Table 4.1.: Computations in Iteration 1.

Data Point	distance to $c_1(3,9)$	distance to $c_2(4,2)$	Cluster
$x_1(3,9)$	0.00	7.07	C_1
$x_2(2,5)$	4.12	3.61	C_2
$x_3(5,8)$	2.24	6.08	C_1
$x_4(7,5)$	5.66	4.24	C_2
$x_5(4,2)$	7.07	0.00	C_2

Now, the centers have to be recalculated as the mean of all assigned data points. The result are the updated center points $c_1(4,8.5)$ and $c_2(4.33,4)$. These are used to compute the distances again for each data point in Iteration 2 as shown in Table 4.2. As the center points have changed, the distances to the centers are different and data points could be reassigned to other clusters. In our made-up example all data points are assigned to the same clusters as after Iteration 1, i.e. no assignment changed. Therefore the algorithm converges and we have found the final result. Figure 4.1 depicts the two clusters: Cluster C_1 with the center point c_1 is marked by blue data points and is positioned on the top of the chart, while Cluster C_2 with the center point c_2 is a broader cluster from the middle of the chart to the bottom.

In this section a made-up example showed the phases of the k-Means algorithm in depth. These fundamentals help us to get a better understanding in Chapter 6, when we discuss the implementation details of k-Means as HyPer operator.

Table 4.2.: Computations in Iteration 2.

Data Point	distance to $c_1(4, 8.5)$	distance to $c_2(4.33, 4)$	Cluster
$x_1(3, 9)$	1.12	5.17	C_1
$x_2(2, 5)$	4.03	2.54	C_2
$x_3(5, 8)$	1.12	4.06	C_1
$x_4(7, 5)$	4.61	2.85	C_2
$x_5(4, 2)$	6.50	2.02	C_2
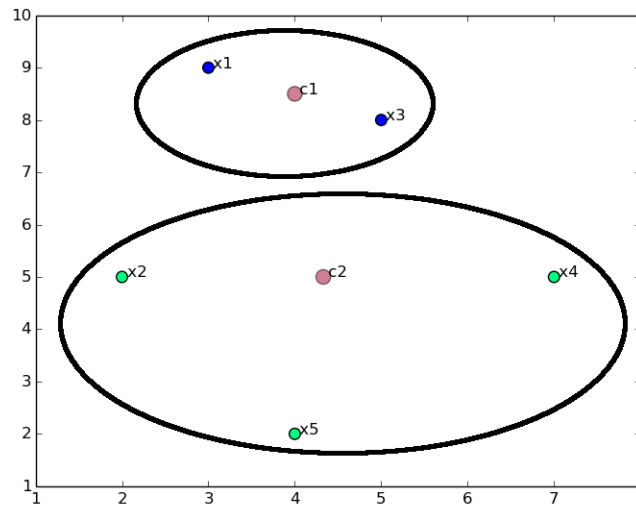
Figure 4.1.: K-Means Example Clustering.

4.4. The k-Means++ Initialization Strategy

Even though the Lloyd algorithm provides good results in practice, the resulting clustering can be arbitrarily bad for some data sets. Particularly, the random choosing of the initial cluster centers can lead to a bad grouping which cannot be changed during the clustering process. Therefore, [AV07] proposes a variation to the k-Means initialization strategy, called k-Means++. The centers are still chosen randomly from the data points, but the data points are weighted according to their distance from the closest, already chosen center. Hence, the probability of choosing data points as center that are far away from each other increases, leading to enhancements in speed and accuracy of the clustering.

Formally, k-Means++ can be described the following: Let $D(x)$ be the shortest distance

from a data point to the closest, already picked center.

1. Take the first center c_1, chosen uniformly at random from X.

2. The next center c_i is chosen from $x \in X$ with the probability

$$\frac{D(x)^2}{\Sigma_{x \in X} D(x)^2}.$$

   This weighting is called the D^2 weighting.

3. Repeat Step 2 until all k centers are selected.

4. Continue as in the standard k-Means algorithm.

Many clustering tools such as Weka and Julia are implementing a k-Means++ initialization strategy, therefore we will also provide this enhancement for the k-Means operator implementation in HyPer and discuss it in Section 6.4.3.

5. Related Systems

In this chapter we present state-of-the-art data mining software, frameworks and languages, their main characteristics and areas of application. We try to give a broad overview starting with research tools and dynamic languages, big data solutions, software systems that strongly integrate with existing databases, and finally database processing of data mining algorithms. For all presented tools we compare their objectives with HyPer's.

5.1. Research Frameworks

In this section we discuss two data mining tools - Weka [Hal+09] and ELKI [AKZ08] - that are strongly originated in the research community, both developed at universities. Additionally, Weka is often used for teaching data mining techniques. Nevertheless, both tools, in particular the more mature Weka, are also used in practice by industrial scientists.

Weka is a data mining "workbench" developed at the University of Waikato that gained big attention in the machine learning and data mining community over the past decade. It provides a large collection of algorithms for all aspects of data mining such as data preprocessing, classification, regression, clustering, association rule mining and visualization into one framework. Weka is written in Java and has with the "Weka Explorer" a built-in GUI, but can also be executed directly within Java programs or from the command line.

Weka's default file format is the arff format, an ASCII text file that describes instances and its attributes. In contrast to csv files, arff files can be read incrementally by Weka, because the header of an arff file already determines whether a column is numeric or nominal. In contrast, all instances have to be inspected when working with csv files. Therefore, arff files provide performance improvements and high accuracy about the attributes' data type.

ELKI (Environment for Developing KDD-Applications Supported by Index Structures) is a relatively new data mining framework developed at the Ludwigs-Maximilians-Universität München since 2009 with an even stronger focus on research. It's main objective is to enable a fair comparison of data mining algorithms based on experimental evaluation. In the strongly evolving field of data mining, many algorithms

are released every year, often never seen again once the paper is published. It is common that the comparison of algorithms is based on experimental evaluation, which is not always comprehensible and reproducible. Moreover, the used code is not always published or properly documented. ELKI addresses these challenges and provides an implementation framework for new data mining algorithms, leading to a better comparability among them and therefore to a fairer evaluation of the newly proposed algorithms.

As Weka, ELKI is written in Java and can be either used via graphical user interface, in Java programs or the command line. A major strength of the framework is that it is able to read arbitrary data types and the support of any distance or similarity measure. Therefore, all kinds of complex data types can be integrated into existing algorithms, merely by providing a distance function for the given data type. ELKI also encourages the use of index structures to achieve performance gains when working with high dimensional data sets.

In contrast to HyPer, Weka's and ELKI's main focus is not on performance but to improve research and teaching in the data mining field. In particular, ELKI's main goal is it to make newly proposed algorithms more comparable. Therefore, both tools are open for extensions from other research groups. HyPer does not intend to be a general research and teaching tool, even though it is used for internal research at the TUM and might be used for teaching data mining algorithms as well. However, HyPer's main objective is to be a standalone high performance database system with data mining capabilities. Instead of working with csv or arff files, HyPer's data mining operates directly in the database engine, implemented in high performance C++ code.

5.2. Dynamic Languages for Statistical Computing

While Weka and ELKI provide an excellent suite for data mining, data scientists often prefer a more dynamic language over sometimes bulky Java code. In particular for exploratory data analysis, dynamic scripting languages are very popular to get a first hands-on-feeling of the given data. Special purpose languages exist, with a syntax and built-in data structures that make data analysis tasks fast and concise. This section presents the R [R D08] language which is popular for decades, the previously announced Julia [Bez+12] language and the Python ecosystem SciPy [J+01]. R and Julia are both special purpose languages designed for data analysis, while Python is a general purpose programming language. However, the SciPy ecosystem provides excellent tools for data mining with a similar syntax and data structures as in R and Julia.

R is a mature language for statistical computing and data analysis with a rich set of

graphical techniques that makes it often researcher's number one tool for creating graphics for publications. Since R is a programming language most of the R libraries are written in R itself. However, C, C++ and Fortran code can be linked and called at run time and is often used for computationally intensive tasks. For advanced usage it is also possible to manipulate R objects directly via C code.

Instead of just being a programming language, R depicts itself as an environment for statistical techniques: It is highly extensible via packages and provides many data mining algorithms, available via the CRAN[1] network. RStudio[2] offers an IDE for R programmers, and software products like JGR[3] provide a GUI for R programs.

Thanks to the underlying C, C++ and Fortran libraries, R can do basic computations like matrix math very efficient and fast. However, the R language interpreter is rather slow, which discourages writing large libraries or complex abstractions in the R languages itself. Instead, the described low-level languages are used.

Julia is a modern dynamic programming language for scientific computing, designed to address some of these concerns. As R, Julia is a programming language, written itself mainly in Julia, using C and Fortran for computationally expensive tasks. In contrast to R, Julia uses an LLVM-based JIT compilation for interpreting its own language. This results in stunning performance results: While the running time for basic language features such as matrix and vector computation does not differ to R, since both are using low-level languages, implementations in the Julia language are much faster and often comparable to C[4]. The same compilation technique is used in HyPer, therefore it will be interesting to compare both tools.

For data analysis, external packages are available providing state-of-the-art data mining algorithms. However, it cannot compete with R's CRAN, one of the most impressive collections of statistical libraries available anywhere. Since Julia is a young language with a very active community and a growing popularity, we can expect many more algorithms and tools in the future.

SciPy is a Python-based ecosystem enhancing the Python language into a data analysis environment. It's core packages are NumPy, SciPy, matplotlib and pandas. The NumPy library provides matrix data structures and efficient computations on top of it, and tools for integrating C, C++ and Fortran code. SciPy includes a very large collection of numerical, statistical, and optimization algorithms; pandas provides R-style Data Frame objects on top of NumPy data structures for fast computation, and matplotlib [Hun07] is a Python 2D plotting library, also used in Julia.

At the same time, Python has already a huge number of well-known libraries, also

---

[1] http://cran.r-project.org/, 15.03.2015.

[2] http://www.rstudio.com/, 15.03.2015.

[3] http://www.rforge.net/JGR/, 15.03.2015.

[4] http://julialang.org/benchmarks/, 15.03.2015.

usable for data analysis. Beautiful Soup[5] is for example a great library to analyze html or xml data. Together with the Python community and other libraries it makes Python a attractive tool for data analysis.

HyPer shares some objectives with the presented tools of this section: A simple dynamic language with data mining supporting data structures and plotting functionalities are long-term goals of computational data mining in HyPer. Furthermore, all tools are using low-level Fortran, C or C++ code for computationally intensive operations, which is a performance improvement over Weka and ELKI, both written in Java. However, as Weka and ELKI, a data extraction, transformation and loading (ETL) is necessary to use the data sets with the presented tools in this section. This makes real-time data mining almost impossible - in contrast, HyPer executes data mining algorithms directly on the running database.

5.3. Big Data Platforms

The presented tools in the previous sections are great when working with megabytes and gigabytes of data, which is very common for data analysis: Often, data scientists spend much time on working on a small sample of a larger set, aggregated results, or simply on a small data set. However, as data size grows big data frameworks like Hadoop are receiving a lot of attention: They provide an affordable, reliable and ubiquitous way to spread computation over tens or hundreds of CPUs on commodity hardware. In this section we discuss the most important trends of the big data community for data mining, in particular Apache Hadoop [Shv+10a], Mahout [Apa] and Spark [Zah+10].

Apache Hadoop is an open-source software for reliable, scalable, distributed storage and processing of large data sets across large clusters of computers. The heart of Hadoop is the Hadoop Distributed File System (HDFS) and Hadoop MapReduce, a simple programming model for distributed processing.

The Hadoop distributed file system (HDFS) [Shv+10b] is a distributed, scalable, and portable file system written in Java for the Hadoop framework. It is highly fault-tolerant and provides high throughput access to application data. HDFS stores large files as blocks replicated across multiple machines.

MapReduce is a programming model originally developed by Google [DG08] for processing large amounts of data in parallel on large clusters of commodity hardware. A MapReduce job splits the input data into chunks processed by the map tasks in parallel. The output of the map tasks is then the input of the reduce tasks which merges the intermediate results. Hadoop MapReduce provides an open-source software framework

---

[5]http://www.crummy.com/software/BeautifulSoup/, 15.03.2015.

for writing MapReduce jobs in a reliable, fault-tolerant manner. Several algorithms have been implemented on the MapReduce programming model, e.g. a parellel k-Means algorithm [ZMH09].

Numerous Apache Software Foundation projects are built on top of Hadoop, such as Hive [Thu+09] and Cassandra [LM10]. For data mining, Apache Mahout [Apa] is certainly the most important project: It is a library of scalable machine learning and data mining algorithms, implemented on top of Apache Hadoop and using the MapReduce paradigm. For data stored on the Hadoop Distributed File System, Mahout provides the data science tools to automatically find meaningful patterns in those big data sets: The Apache Mahout project aims to make it faster and easier to get information out of large data sets. Therefore, Mahout provides the implementation of various data mining algorithms for Hadoop, executable in local mode (if MapReduce is not applicable) or in distributed mode.

In 2014, the Mahout community decided to move its code base onto a more modern data processing system that offers a richer programming model and more efficient execution than Hadoop MapReduce: Apache Spark. Apache Spark is a data analytics cluster computing framework able to work on top of the Hadoop Distributed File System, but also with data from Hive, Cassandra or any other Hadoop input format. In contrast to Hadoop's MapReduce, Spark's richer programming model leads to tremendous performance gains for some applications. Spark also provides in memory cluster computing, making it well-suited for data mining algorithms.

Even though the presented big data solutions show impressive results for large data sets, setting up large clusters of commodity hardware is a complex process. In contrast, HyPer's database and the data mining algorithms run on a single, modern computer. With the growing number of cores and multi-threading capabilities of modern hardware, we also expect good performance for data mining on large data sets with HyPer. Furthermore, big data solutions also require some kind of ETL process, for example to get the data into the HDFS, which is obsolete for HyPer.

5.4. Middleware Tools

While all of the presented environments are frequently used by data scientists, they demonstrate one decisive drawback: Before executing data mining algorithms, the data first has to be fetched from the database and transformed into a format readable by these tools. Therefore it is an enhancement to bring the algorithms closer to the database. In this section we present MADlib [Hel+12] and RapidMiner [Rap14], two tools that try to fill this gap.

MADlib is an open-source library providing a suite of SQL-based algorithms for ma-

chine learning, data mining and statistics, that can run within a database engine. Therefore, it is not necessary anymore to import or export data with an ETL process from the database to the respective data mining tool. The analytical algorithms of MADlib can be installed and executed within a relational data base engine as long the engine supports extensible SQL. So far, MADlib works with PostgresSQL[6], Pivotal Greenplum[7] and Pivotal HAWQ[8].

MADlib's main algorithms are written in declarative SQL statements using extensible SQL; higher level tasks such as iterations and specific program structures are implemented in Python code. This code drives the algorithms and invokes data intensive computations directly in the database.

RapidMiner is another tool for predictive analysis with a strong focus on visual development. It's main goals are low entry barriers and quick speed up. Therefore, it's main audience are not only data scientists and IT-specialists, but also the business department and stakeholders. To achieve this, RapidMiner let end users to design data analysis routines with a graphical user interface.

RapidMiner provides three different analytical engines for data mining. By default, all algorithms and computations are executed in memory. Since random access is often necessary for data mining algorithms this is the fastest approach for medium size data sets. In contrast, in database mining brings the algorithms into the database, therefore the loading phase of the memory processing gets redundant. This solution is offered for different database systems utilizing database functionality. Finally, in Hadoop computation allows to combine the user interface of RapidMiner with the workflows of Hadoop clusters to analyze large data sets.

Interestingly, the execution of algorithms in the database is not designed for performance, but for high data locality. RapidMiner's performance is the best when the data is first loaded and executed in the in memory engine. Data mining algorithms in HyPer can benefit from both: High data locality and high performance. Even though both tools objective is to bring data mining algorithms to the database, often "glue" code is necessary to make data mining algorithms work, often affecting the performance of the algorithm. Furthermore, not all database systems are supported by MADlib and RapidMiner.

---

[6]http://www.postgresql.org/, 15.03.2015.

[7]http://pivotal.io/de/big-data/pivotal-greenplum-database, 15.03.2015.

[8]http://pivotal.io/de/big-data/pivotal-hawq, 15.03.2015.

5.5. Data Mining in Databases

Software like MADlib and RapidMiner often combine SQL operators with high-level programming constructs. Since this is not ideal for the performance of the algorithm, database vendors such as Oracle and SAP are going one step further and providing their databases with more and more statistical and data mining functionalities using the existing SQL syntax and graphical user interfaces.

SAP HANA [Fär+12] is SAP's main memory database. Its functionality is very similar to HyPer's: It combines transactional processing (OLTP) and analytical processing (OLAP) into one system without the need of a database and an additional datawarehouse. SAP PAL [SAP14] is an extension for HANA to provide various data mining algorithms. The idea is similar to HyPer's: Instead of importing and exporting data to other, external tools, the best place to perform data mining algorithms is directly in the database. Therefore, PAL's complex analytical computations can be performed very efficient and fast in memory. The transfer time of large tables from the database to the application becomes redundant and calculations are much more inexpensive.

As one of the market leaders for traditional, disk-based databases, Oracle provides a lot of possibilities for data mining and knowledge discovery its databases, such as the Oracle Analytical SQL Features and Functions[9] and the Oracle Statistical Functions[10]. The Oracle Analytical SQL Features and Functions are a suite to extend the already existing SQL syntax by a wider range of analytical features such as rankings, windowing and reporting aggregates. Oracle Statistical Functions enhance the normal functionality by statistics such as descriptive statistics, hypothesis testing, correlations analysis, cross tabs and the analysis of variance.

On top of this ranks the Oracle Advanced Analytics Options [Ora15]. It provides techniques for state-of-the-art data mining technologies by implementing in database algorithms in the SQL language and by integrating R algorithms. Oracle Data Mining (ODM) provides algorithms as native SQL for high performance. As previously, an importing or exporting of data to other tools becomes redundant. Users can either use the Oracle Data Miner[11], a work flow based GUI tool, or a SQL API. Additionally, Oracle R Enterprise (ORE)[12] integrates the already presented R language for statistical computing with the database. Own R packages and extensions can be implemented and executed in the database without additional data loading.

---

[9]http://www.oracle.com/technetwork/database/bi-datawarehousing/sql-analytics-index-1984365.html, 15.03.2015.

[10]http://www.oracle.com/technetwork/middleware/bi-foundation/index-092760.html, 15.03.2015.

[11]http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html, 15.03.2015.

[12]http://www.oracle.com/technetwork/database/database-technologies/r/r-enterprise/overview/index.html, 15.03.2015.

Even though data mining in SAP Hana and the Oracle database is closely related to HyPer's objective, such as data locality, high performance computing in the database and real-time data analysis, our approach is a slightly different one: SAP HANA provides with PAL data mining capabilities strongly related to the SQL language; Oracle offers data mining algorithms invoked by SQL or by graphical user interface, and enables the execution of R algorithms directly on the database. In contrast, HyPer provides a system with high performance data mining algorithms, that run directly on the database, but can be invoked in a simple dynamic language similar to R, Julia and Python. Therefore, data scientists can benefit from the performance of in database processing of data mining algorithms and the simplicity of well-known functional languages.

In this chapter we gave a broad overview over state-of-the-art data mining software, frameworks and languages, and their objectives in contrast to HyPer. A subset of these tools will be used in the experiments in Chapter 7 to compare their performance against the performance of the HyPer k-Means operator, which implementation will be described in the next chapter.

6. Implementation of k-Means in HyPer

In this chapter we present the implementation of the k-Means algorithm as a HyPer operator. First we discuss the general implementation of operators in HyPer, the used programming model and the LLVM code generation. Then, the constraints and requirements of the k-Means implementation are introduced. Next, the implementation details are depicted including data materialization, a C++ driven version and an LLVM driven version. Additionally to the random initialization strategy the k-Means++ initialization strategy is implemented as described in Section 4.4. Finally, a parallel realization of the k-Means operator is discussed.

6.1. HyPer Operator Fundamentals

In this section we discuss the general implementation of operators in HyPer and the involved code generation. With this programming model in mind we can afterwards show how the k-Means operator is implemented in HyPer. The argumentation and results of this section are based on [Neu11] and [Neu14].

6.1.1. The Consume Produce Programming Model

As all data resides in memory, main memory databases' query performance is much more dependent on the CPU cost of the query processing than on I/O cost as in traditional systems. Therefore, query processing has to be reinvented to achieve optimal performance. Before a query is executed, most database systems translate a query into an algebraic expression and start evaluating and executing this algebraic plan. Traditionally, this is done using the iterator model [Lor74]: Each physical operator produces a tuple stream and iterates to the next tuple by calling the `next` function of the operator. The iterator model works well for I/O dominated, traditional databases, where CPU consumption is not a limiting factor. However, for main memory databases, this is not perfect: First, `next` is called many times, once for every single tuple for each intermediate and final result. This `next` call is usually a virtual call or a call via function pointer which makes it more expensive than a regular call and reduces branch prediction of modern CPUs. Moreover, the iterator model results often in poor code locality and complex bookkeeping. This can be seen from the functionality of the

table scan operator: As tuples are generated one after the other, the operator has to remember where in the compressed stream the current tuple was and has to jump back when asked for the next one.

In order to resolve these issues, Neumann [Neu11] proposes a new query compilation strategy for main memory databases: Instead of the operator centric approach of the iterator model, processing is data centric. Therefore data can be kept in the CPU registers as long as possible, while the boundaries between the operators get more and more blurred. That means that each code fragment performs all actions on the given data, until the result has to be materialized, i.e. data is taken out of the registers. A code structure like this generates almost optimal assembly code, since all the relevant instructions for the given data are generated, and therefore the data can be kept in the CPU registers.

HyPer uses a simple programming model for its operators in order to make compilation as efficient as possible, while writing code remains understandable and maintainable for the developer: Each operator implements two functions, a `consume` and a `produce` function. The `produce` function computes the result tuples of an operator, which are then pushed to the next operator by calling its `consume` function. The next operator works the same way, after getting data in by a `consume` call of the predecessor, it produces result tuples by calling its own `produce` function. To wrap it up, each operator gets its own `consume` function called by its predecessor, calls its own `produce` function to compute the results and calls then the `consume` function of its successor. This process is shown in the sequence diagram in Figure 6.1.
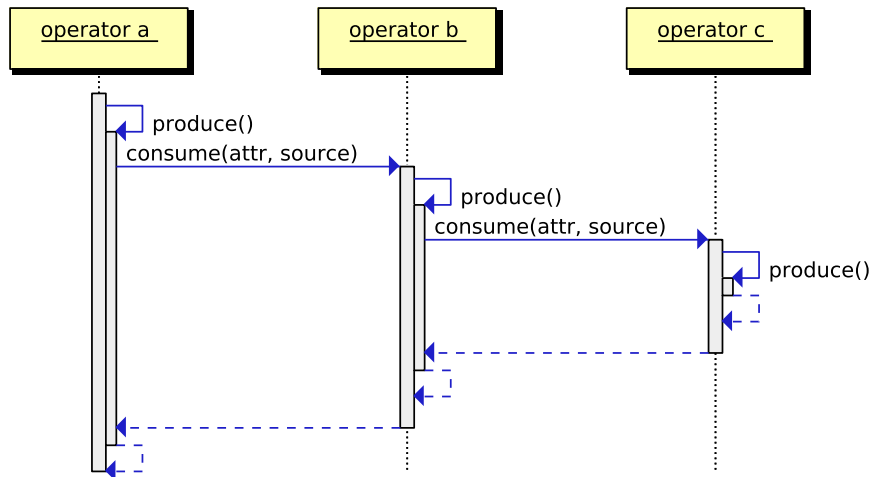

Figure 6.1.: Consume Produce Sequence Diagram.

Therefore, this programming model pushes data towards the next operator, instead

of pulling the data. This results in a much better code and data locality. Tuples are pushed from one operator to the next, therefore operators benefit from keeping the data in the CPU registers, which allows very cheap and efficient computation. Only for materializing data memory has to be accessed. Additionally, small code fragments are used to handle large amounts of data in tight loops leading to good code locality and therefore high performance.

These aspects make the execution of operators very fast. Unfortunately, not all operators can be implemented in a way that one tuple gets fully processed before the next one is loaded into the CPU registers. Instead, tuples have to be taken out of the registers and are materialized in main memory for some operators. Those operators break the operator pipeline, therefore we call them *pipeline breakers*. If an operator materializes all incoming tuples before continuing processing, we call it a *full pipeline breaker*. An example is a join operator: One side of the two join relations has to be materialized in main memory, while the other relation is scanned and probed against the materialized data for finding join partners. As the join operator materializes one relation it is a *pipeline breaker*.

Finally, it is important to note that the `consume produce` model is just an abstraction layer for the programmer. This abstraction layer is used by the code generation to compile assembly code - within the generated assembly code there is no `consume produce` present anymore.

6.1.2. LLVM Code Compilation

In this section we discuss the query compilation in greater detail. As in traditional systems, queries are compiled by parsing the query, translating it into algebra and optimizing it. In contrast to a traditional system, the physical algebra is not directly executed, but compiled by the code generation into an imperative program. For generating this imperative program the previously described `consume produce` model is used.

Taking advantage of the structure of the `consume produce` model, queries are compiled into native machine code using the LLVM compiler framework [LA04]. LLVM can generate portable assembler code which can be executed directly using an optimizing JIT compiler. With LLVM HyPer uses a very robust assembly code generation, e.g. pitfalls like register allocation are hidden by LLVM. Therefore the assembly code generation is very convenient compared to other compiler frameworks. Furthermore, only the LLVM JIT compiler translates the code into machine dependent code, leading to portable code across computer architectures. Since the LLVM assembler is strongly typed, many bugs can be caught in advance, in contrast to a textual C++ code generation. Furthermore, LLVM produces highly optimized, extremely fast machine code and outperforms in

some cases even hand-written code as the assembly language allows code optimization, hardware improvements and other tricks, that are hard to do in a language like C++. All this requires usually only a few milliseconds of compilation time, which is a crucial criterion for database operators with a strong focus on performance.

Additionally, LLVM code is perfectly able to interact with C++, the main language of the HyPer database. Even though LLVM code is robust and convenient to write compared to common assembler code, it is still more painful than writing code in a high-level language such as C++. An interaction of the two languages allows us to implement complex structures in C++ and to reuse existing database logic such as index structures.

Therefore algorithms can be written in C++ and connected together by LLVM code, where the C++ code is pre-compiled and the LLVM code is compiled at runtime dynamically. An example is the Sort operator: The `compare` function, comparing two tuples by the rules of a sort query is dynamically generated in LLVM code, depending on the schema of the database. As actual sort function, the built-in C++ `sort` can be used. This is a great example of the mixed execution model using C++ and LLVM code in one operator.

However, even though C++ and LLVM can both be used implementing an operator, LLVM code is dominant and C++ code should be seen as convenience. For performance reasons it is important that the code executed for most of the tuples is generated code, even though calling C++ from time to time is acceptable. As already mentioned, staying in LLVM allows us to keep the data in the CPU registers and is therefore the preferable way of executing a query. Calling an external function spills all registers to memory, which can be a bottleneck when doing this for all tuples in a big data set.

In conclusion, we have presented the HyPer `consume produce` programming model which allows operators to push data towards the next operator in an efficient manner. Furthermore, the LLVM compiler framework with its benefits is introduced used for code generation and interacting with C++ code at runtime. This cooperation of LLVM and C++ code regarding programmer friendliness and execution time is one of the dominant patterns in Section 6.4, where two alternative implementations of the k-Means operator are presented and discussed.

6.2. Requirements and Constraints

Before discussing the technical implementation details of the k-Means operator, we introduce the requirements and constraints. Obviously, the k-Means algorithm must be implemented as a HyPer operator. That means, the `consume produce` programming model and the code generation with C++ and LLVM have to be used. The advantage of

implementing k-Means as an operator is that it can be used in combination with existing SQL operators, useful for data mining as well, such as grouping and aggregation.

Regarding the functional aspects, the operator must be able to be executed in serial and in parallel, to make use of the computing power of modern database hardware.

As input parameter, the user can specify the number of maximum iterations of the algorithm. If this parameter is omitted, the algorithm runs until convergence. For big data sets this is often a performance problem when only a few data points are changing, but still all the distances have to be computed all over again. Often the result is already accurate enough after a specific number of iterations. Apart from an advantage regarding the running time it is also useful to specify the number of iterations for experiments and evaluation, as we see in Chapter 7.

Further parameters are the initialization strategy and a verbose option. As initialization strategy, the user can select between random initialization and the k-Means++ initialization, as described in Section 4.4. As default output, an additional column is added to each data row presenting the cluster identifier of the data tuple as an integer. If the verbose option is active, statistics of the run are printed to the console. This information contains the number of iterations, the final center coordinates, the number of assigned data points per center and the squared error sum.

The number of iterations is crucial for comparing the running time of different k-Means algorithms in a fair manner. It is possible that one execution of the algorithm converges after three iterations, while another one converges after seven iterations, since the number of iterations is non-deterministic due to the random initialization strategy. For fair evaluation, the time per iterations is therefore a good quality measurement and will be used in the evaluation in Chapter 7.

6.3. Data Materialization

In this section we discuss the data materialization of the k-Means operator. As already mentioned, materialization is the process of taking incoming tuples out of the CPU registers and write them into memory. Since this decreases the performance of the database, HyPer tries to avoid it, and instead keeps the data in the pipeline as long as possible, i.e. until a *pipeline breaker* occurs. Unfortunately, k-Means is a *pipeline breaker*, e.g. center tuples have to be compared with all data tuples to find the minimum distance between them. Therefore, we have to put all tuples into memory. This is depicted in Figure 6.2: Each incoming tuple will be written into main memory when entering the `consume` function.

This is implemented by using a combination of LLVM and C++ code. In HyPer, LLVM code resides in the `compile time system (cts)` while C++ resides in the `runtime`
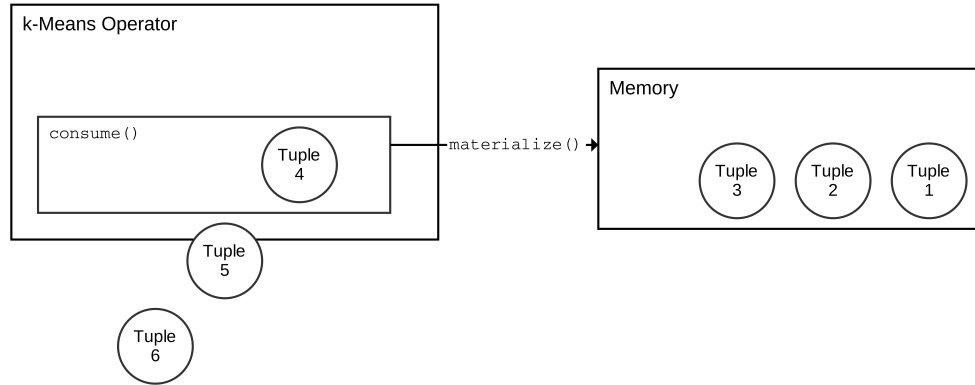
Figure 6.2.: Data Materialization of Incoming Tuples.

`system (rts)`. Since the `compile time system` is the entry point of an operator, the `consume` function is called and generates code for each tuple. This generated code is materializing the incoming tuples. A pointer to the materialized chunk of memory is then stored in a C++ vector in the `runtime system`. Figure 6.3 depicts this process: Each tuple is materialized into memory by the `cts` and can be referenced by a pointer stored in a vector in the `rts`. This vector can later be used to loop through the entire data set: First by looping through the vector in the `rts`, getting the pointer to the memory location, which can then be used to load the tuple from memory into the CPU registers in the `cts`.

For k-Means it is not enough to store only the data tuples. We also need to reserve memory space for the centers. We do this by materializing the first `k` tuples of the data set two times: Once as storage for data tuples and once as storage for center tuples. Therefore we have two vectors in the `runtime system`: One for data tuples of length `n`, and one for centers of length `k`. This data materialization for k-Means is depicted in Figure 6.4.

At this point of the process, all data tuples and centers are materialized and an additional field has been added to all of them: A cluster identifier of type integer. For the center tuples, this field stores the center identifier, which goes from 0 to $k - 1$. For data tuples, this field specifies to which center and therefore to which cluster a data tuple belongs to. Initially, all data tuples are assigned to the center with identifier 0.

Another difference is the cast of data types between data and center tuples. While the data types of data tuples are specified by the table schema, the data types of the center tuples are determined differently since the center is the mean of all the data tuples assigned to that center. Therefore, an integer data value is stored as a numeric data
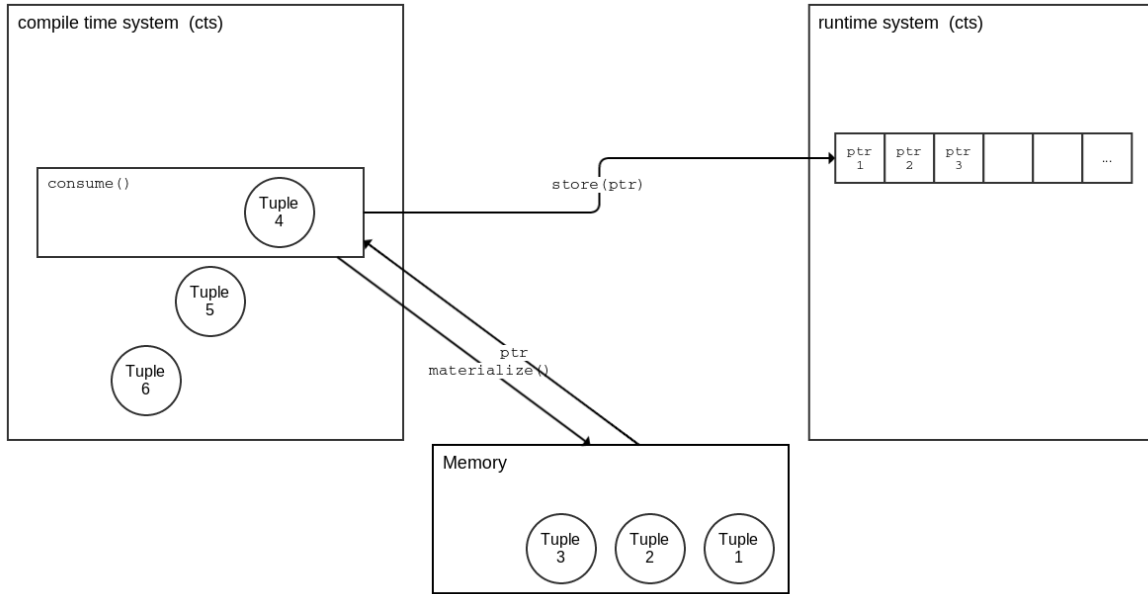
Figure 6.3.: Data Materialization using the Runtime and Compile Time System.

type as center value. Numeric is an exact-value data type for integer and decimal types, and therefore able to store the mean of integer data tuple values. For each data type there exists a rule to cast the respective type from data to center tuple.

6.4. Serial Implementation

After acquiring a basic understanding of HyPer's operator model, in particular about the interaction of dynamically generated LLVM code and pre-compiled C++ code, we can discuss the actual implementation of a serial k-Means algorithm. The most interesting part of implementing a data mining operator like k-Means in HyPer is the decision of the design of the algorithm, i.e. which parts of the code should reside in the `compile time system` and which parts in the `runtime system`.
This question is not trivial because there are no strict rules and several possibilities. For example, a dynamic generation of code works best for comparing tuples with each other as in the sort operator, or the computation of a distance in the k-Means operator. This code has to handle different data types depending on the table schema and therefore generated code is preferable. For other parts, like the implementation of a sort function or the combination of loops in k-Means, it is not so obvious where to put the code.
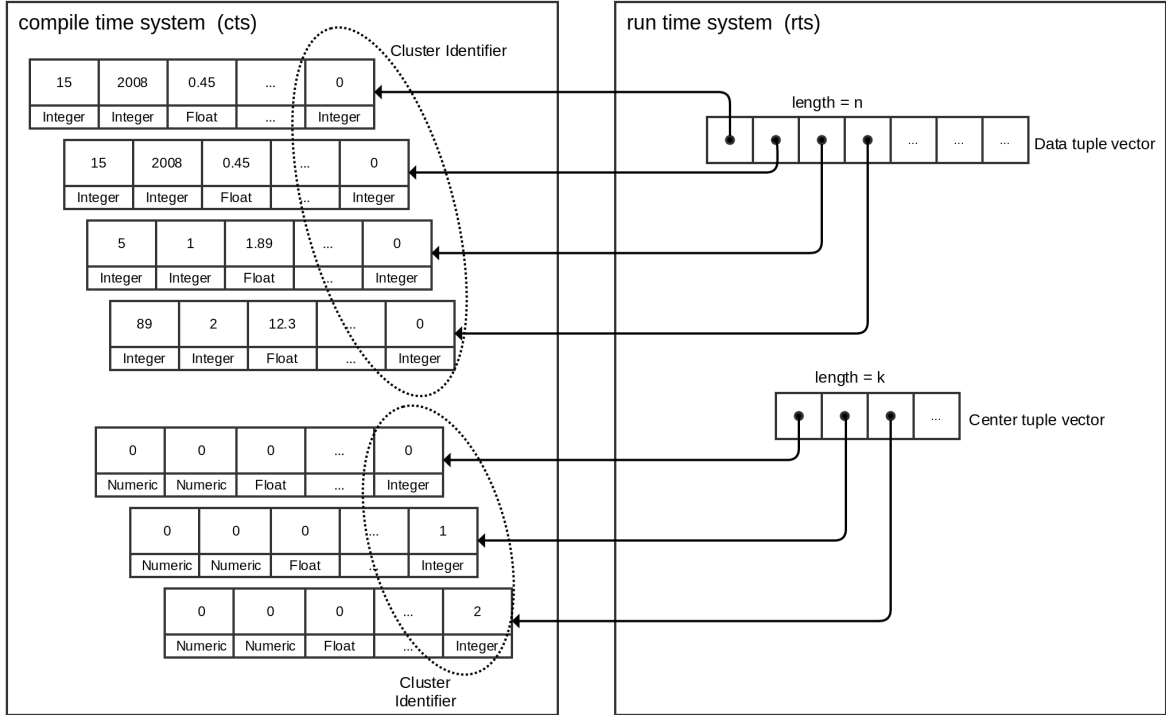
Figure 6.4.: The k-Means Data Materialization in Detail.

In this work we present two different ways of implementing a serial k-Means operator in HyPer, first by implementing the algorithm in C++ with only a few generated LLVM functions, e.g. to compute the distance. Then, a system is presented implementing k-Means almost entirely in LLVM. Only small parts, like initializing the random centers are implemented in C++. This implementation benefits of generated, compact code and we expect performance gains over the C++ driven implementation.

6.4.1. A C++ driven Implementation

As first implementation approach we present a C++ driven version. The term C++ driven is maybe misleading, since all operators start their main computation after invoking their `consume` function, which is an LLVM generated function. Even though the operator starts working in the `compile time system`, in this implementation, `cts` calls the k-Means operator of the `runtime system` and gives the full control to the C++ code, until the k-Means algorithm terminates.

Figure 6.5 shows this interaction in a sequence diagram: The algorithm is invoked in the `compile time system` and calls the k-Means function of the `runtime system`. The entire execution stays now in the `runtime system`, with calls back to the `cts` from time to time, but immediately returns to the `rts` once a function call terminates.

At first, the centers are selected. Let's assume we are using a random initialization strategy. `K` times, a random pointer of the data vector is selected, and its values are copied to the corresponding center tuple. Therefore, the `centerPtr` and the randomly selected `dataPtr` are parameters of a generated `setCenter` function. This function loads both tuples from memory using its corresponding pointers. Next, the data values are casted if necessary, since we have different data types among center tuples and data tuples as described in the previous section, and then copied to the center tuple. Afterwards, the center tuple is written back to memory. This process is repeated until all *k* centers are determined.
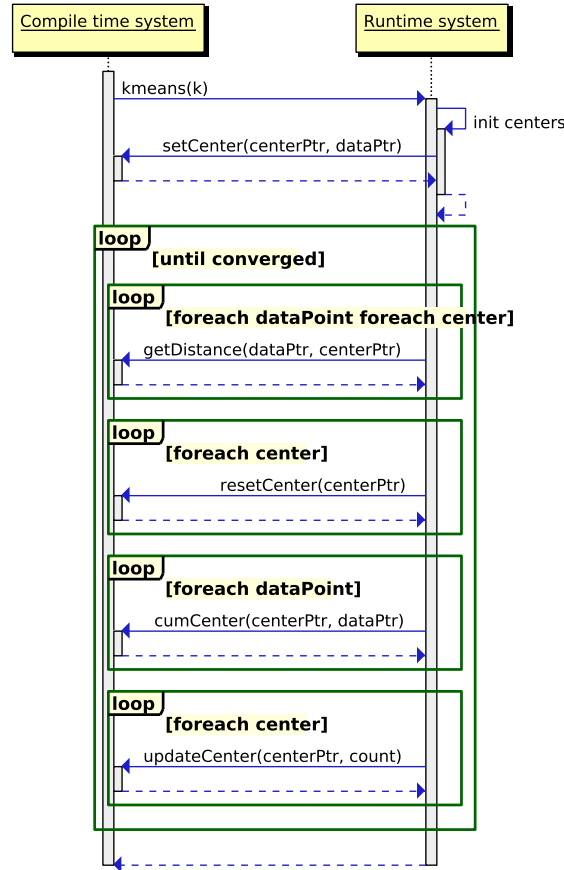

Figure 6.5.: The k-Means Algorithm - C++ driven.

After selecting the initial set of centers, the `runtime system` starts its outer loop, running until k-Means converges. In the outer loop, the `runtime system` calls the `compile time system` to compute the distances between all data points and center points to find the closest center for each data tuple. Therefore, `getDistance` is invoked for each `dataPtr` - `centerPtr` combination. The closest center for each data point is then stored in an `unordered_map` in C++ for efficient lookup.

After finding the closest center for each data tuple, the centers have to be updated. Since the materialization of centers is done in LLVM code, the `runtime system` has to call the `compile time system` again, in fact several times: First, a `resetCenter` function is called for all center pointers to set the center values of the center tuples to zero. Then, the new mean can be computed.

The computation of the mean is done in two steps. First, the data points are accumulated for each center, and then divided by the number of data points belonging to each center. This simple mean computation leads to several `compile time system` calls for our algorithm. For each data point, its values are added to the center tuple it belongs to. Therefore, the `cumCenter` function is called, adding the values of the data point to the center point and also updating the cluster identifier of the data tuple. The `runtime system` keeps track of how many tuples are added to each center. When finished, each center is called again to compute the actual mean of the center dividing by the count of data tuples assigned to the center using the `updateCenter` function. This concludes the first iteration of the algorithm. The next iteration starts then again with the computation of the closest center for the data points. This process continues until the algorithm converges.

Using this implementation approach the main control of the algorithm remains in the `runtime system`. However, this kind of implementation leads to many calls between the `compile time` and the `runtime system`, in total $(n+2) \cdot k + n$ calls per iteration, as Table 6.1 shows. The next section presents an implementation that prevents the algorithm from too many calls between the two systems, which might affect the performance of the operator in a beneficial way.

Table 6.1.: Generated Function Calls per Iteration.

Generated Function	Calls per Iteration
`getDistance`	$k \cdot n$
`resetCenter`	k
`cumCenter`	n
`updateCenter`	k
Total	$k \cdot n + k + n + k = (n+2) \cdot k + n$

6.4.2. An LLVM driven Implementation

As already stated, LLVM code generation encourages the interaction between LLVM and C++ code which is exploited a lot in the C++ driven implementation of k-Means: The algorithm is implemented in C++ and benefits from high-level data structures such as `unordered_maps` and the convenience of the C++ syntax, allowing a quick implementation.

However, the approach presented in the previous section leads to many calls between the `compile time` and the `runtime system`, as shown in Table 6.1, which could possibly affect the performance of the operator. Therefore, a second approach has been explored, implementing the k-Means algorithm almost entirely in LLVM code. Only the random initialization of the center points remains in C++. When thinking about such an implementation, we have to be aware that we cannot use C++ high-level data structures anymore, but have to work with what LLVM gives us.
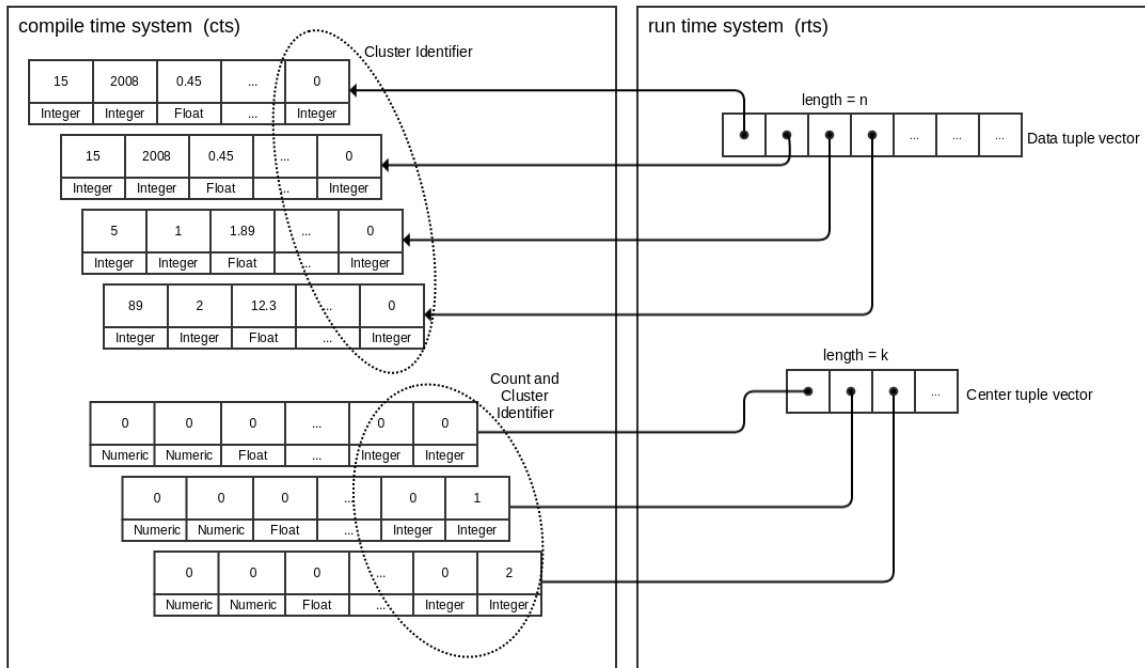

Figure 6.6.: The k-Means Data Materialization - LLVM Implementation.

The only data structure we have used so far in LLVM was a structure to materialize the data and center tuples. In order to keep things simple, we exploit this data structure even further to use it with the LLVM k-Means implementation, without using any other

data structures in addition. So the question is: How do we extend the existing data structure of the `compile time system` to emulate the C++ code we want to omit? This can be done by extending the center data structure when materializing the centers in the `consume` function. As Figure 6.6 shows, an additional field has been added to the center tuple to keep track of the count, determining how many data points are close to this center. With this small modification we can implement our algorithm in LLVM with the use of only one data structure.

The LLVM driven algorithm is depicted in a sequence diagram in Figure 6.7 and shows the indirection of the calls compared to the sequence diagram of the C++ driven approach in Figure 6.5: This time, the LLVM code executes the algorithm, calling C++ code from time to time. There are also no loops around the calls between the LLVM and C++ code, therefore the number of calls is very low.

As in the previous implementation, the algorithm starts with selecting the random centers. This process does not change, but afterwards the control of the algorithm goes back to the `compile time system`. The only information the `compile time system` requires from the `runtime system` is the first pointer and the last pointer of the data vector and of the center vector, respectively. Then, the `compile time system` is ready to execute the k-Means algorithm without any further interaction with the `runtime system`.

Next, the minimum distances are computed between center and data tuples. The distance function was already implemented in LLVM code for the C++ driven approach and the code can be reused, only the loops around the `getDistance` function have to be implemented in LLVM.

The next step is to update the center tuples. The functionality of the `resetCenter` function can be kept the same, as well as the `cumCenter` function. Again, the loops around the functions are now implemented in generated LLVM code. When computing the mean of a center, we have to keep track of the count. For that purpose we are using the additional field of each center tuple created on data materialization. Whenever we are adding data tuples to the corresponding center, we increment the count. When invoking the `updateCenter` function, this count is used to compute the mean using a division.

Even though the difference between the two presented approaches does not seem to be huge since the overall programming concept remains the same, the difference in the code is significant. In particular using LLVM over high-level C++ constructs adds an overhead in the number of code lines. Therefore, the LLVM code in the second approach is harder to understand and to maintain. On the other hand, we are shrinking the number of calls between the C++ and LLVM system from $(n + 2) \cdot k + n$ down to 4. This leads to performance improvements as we see in Chapter 7, where the LLVM driven approach shows better results for all carried out experiments.
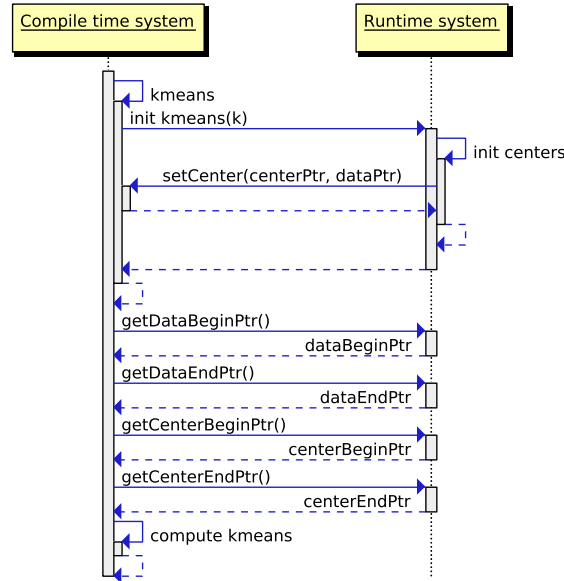
Figure 6.7.: The k-Means Algorithm - LLVM driven.

6.4.3. Initialization Strategy

In the so far discussed implementation details of the k-Means algorithm we assumed to use a simple random initialization strategy. Such a strategy can be implemented in the `runtime system`, since only C++ code is used for finding a random subset of length `k` of the data tuples, which can be implemented using the Fisher-Yates shuffle algorithm [FF48].

However, in Section 4.4 we figured out that one of the most popular variations of k-Means is k-Means++. The k-Means++ algorithm uses an extended initialization strategy by choosing data points as center point with higher probability the further away they are from the already chosen set of center points. Often, this leads to improvements in both speed and quality of the clustering.

For implementation this means that we have to compute the distance from the chosen center points to the data points in the initialization phase too. This can be done very conveniently for the C++ version, as the `getDistance` function is implemented already. Therefore, the main initialization routine can be written in C++, only the `getDistance` function is used as generated function.

For the LLVM version, we still keep the center initialization in the `runtime system` to benefit of high-level C++ program structures. To realize the k-Means++ algorithm, we have to add an explicitly generated `getDistance` function to the LLVM code: Even though the LLVM version computes the distance as well, there is no generated function

anymore, since the distance generation is part of the generated program. Once we have added this function, k-Means++ can be implemented the same way as in the C++ driven approach.

6.5. Parallel k-Means

After presenting two single-threaded implementations of the k-Means algorithm, in this section we show an approach to implement k-Means in HyPer in a parallel way to make use of all cores. Keeping in mind that HyPer is a high-performance database system written in C++, it makes already excessive use of multi-threaded programs, therefore it is only logical to add a parallel version of k-Means too. In the following we use the serial C++ version and modify it to allow parallelism. The C++ version is used over the LLVM version due to higher maintainability, readability and the limited time of this work.

The main bottleneck of the serial implementation is to compute the closest center for each data point: The algorithm has to calculate the distance to each center point which has to be performed in each iteration. This means $k \cdot n$ independent distance computations, which can benefit a lot from parallelism and a computation in independent threads.

When executing a HyPer operator in parallel, the `consume` function is called for chunks of input tuples instead of the entire data set. Consider a system with four threads as shown in Figure 6.8: Each thread consumes one fourth of the data set and materializes the input tuples. Also in the `runtime system`, there are now vectors for each thread, storing the pointers to the input tuples.

After consuming the data the k-Means algorithm continues with choosing the initial center set. For that, we have to break with the parallelism: For a random initialization of the clusters, center tuples are selected from the entire data set randomly. This is even more important for the k-Means++ initialization strategy computing centers regarding a distance function. While we could select random data points for the center just from one data chunk, this is not possible anymore for the advanced logic of the k-Means++ initialization strategy. Therefore, we create a global vector storing pointers to all materialization tuples. Now we can select the global center points using one of our initialization strategies and continue the parallel program sequence afterwards.

After this serial initialization we can continue the parallel execution of the algorithm and compute the distance from each data point to each center point. This expensive computation can be done independently in each data chunk on a separate thread and benefits from the potentials of parallelism. Ideally, this leads to a running time advantage of one fourth of the serial execution, however, due to the overhead of process
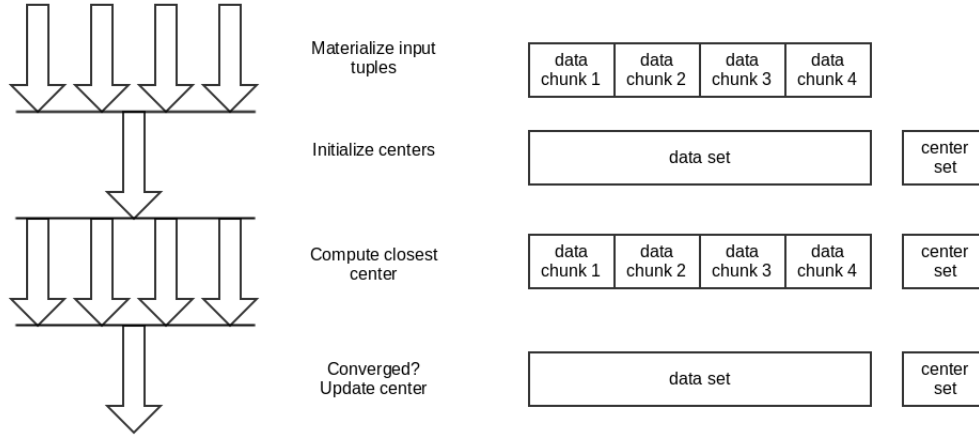
Figure 6.8.: The k-Means Algorithm as Parallel Operator.

creation and management we will not be quite able to attain such high improvements, as the experiments in Chapter 7 show.

After this parallel computation, we jump back to a serial execution mode and check for each thread if the program has converged. Only if all threads have converged, we can terminate the algorithm and output the result of the clustering. Otherwise, we update the center points by the newly assigned data points. In the first version of this parallel implementation, this step is done in serial as already presented in Section 6.4. However, here is potential for parallelizing the program even more, which is a goal for future implementations as we see in Chapter 8.

In the evaluation chapter we compare the parallel k-Means algorithm with the two serial implementations. Even though only one part of the algorithm is parallelized and with the overhead of introducing the parallelism in mind, we expect huge performance gains since the parallelized code part contains many independent computations well-suited for multi-threaded execution.

7. Evaluation

In this chapter we provide an extensive experimental evaluation of the HyPer k-Means operator and compare the different HyPer implementations to each other as well as to some of the state-of-the-art clustering tools presented in Chapter 5. All experiments have been performed on a workstation equipped with sixteen 2.93 GHz CPUs and 64 MB of main memory.

7.1. Data Sets

As objective comparison criterion the execution time per iteration is measured in the following experiments, as described in 6.2. Therefore the different technologies can be compared in a fair way and random initialization and internal improvements for faster convergence do not affect the outcome. For the experiments a real world data set is used and four synthetic data sets have been generated[1]. The synthetic data sets are chosen in a way that both the number of instances and the number of dimensions are varying, covering many real world cases and also allow a reasonable comparison between the data sets. The used data sets are shown in Table 7.1.

Table 7.1.: The Data Sets.

	Instances	Dimensions	Size (byte)	Size (Gigabyte)
3D Network	434,874	4	20,673,913	0.019
High Dimensional	50,000	50	10,000,000	0.009
Medium Size	15M	4	240,000,000	0.22
Medium Size HD	15M	50	3,000,000,000	2.79
Large Size	150M	10	6,000,000,000	5.59

- 3D Network: The first data set contains the 3D spatial road network data of a road network in North Jutland, Denmark [KYJ13]. It is a real world data set available at the UCI Machine Learning Repository [FA10] and consists of 434,874 data points

---

[1]The data set generation script can be found in Appendix A.

in four dimensions: The id of the road segment, the latitude, the longitude and the altitude of the segment. The id is a large integer, the other dimensions are floating point numbers. The size of the data set is 19.72 MB, therefore it is a rather small data set.

- High Dimensional: As second data set a synthetic data set with 50,000 data points has been generated. The data set consists of 50 dimensions and is used to resemble a rather small but high dimensional data set. All dimensions are floating point numbers. The size of the data set is 9.50 MB, hence an even smaller then the network data set.

- Medium Size: Next, a larger data set has been created: The medium size data set consists of 15 million data points in four dimensions. The size of the data set is 228.88 MB, therefore almost 12 times larger as the network data set. Data sets with millions of instances are very common numbers for data mining.

- Medium Size High Dimensional (HD): For comparing the previous data set with a high dimensional data set, the medium size high dimensional data set has been generated. As the medium size data set, it consists of 15 million data points, this time in 50 dimension. This leads to a growth from 0.23 GB to 2.79 GB. The same number of instances with different dimensions allows us to evaluate how different algorithms behave when dimensions are growing.

- Large Size: Finally, we generated a even larger data set. The number of instances increased by a factor of 10 compared to the medium data sets resulting in 150 million instances. As number of dimensions, 10 has been chosen as a middle way between the previously chosen 4 and 50 dimensions. This leads to a data set of 5.59 GB.

Additionally to varying the data sets, we also vary the cluster number k from 3, 10 and 20 for each data set for each used tool. Since none of the data sets has a natural optimal clustering, there is no perfect k, therefore a varying k works best for comparing different tools.

7.2. Competitor Systems

As already described in Chapter 5, there exist many software solutions and programming languages for data mining. In this section we look at a subset of the tools implementing the k-Means clustering algorithm we are using to compare with our own k-Means HyPer operator. To make the results comparable in a fair manner, we only

use tools that give enough information about the clustering process, e.g. the number of iterations, the cluster compactness and most importantly, the applied algorithm. Not all tools are implementing the Lloyd algorithm or, as for R, the default algorithm is the Hartigan and Wong algorithm, and the usage of the Lloyd algorithm has to be explicitly specified as an input parameter.

Since k-Means is a non-deterministic algorithm, the number of iterations is the most important criterion to make the results comparable. The running time of k-Means depends almost entirely on the number of iterations the algorithm has to make, which is largely influenced by the non-deterministic initialization. Therefore the running time should be relative to the number of iterations. Unfortunately, neither ELKI's nor Scipy's k-Means implementation provide the number of iterations as output. Therefore, a fair comparison is not possible. Fortunately, Weka, R and Julia provide good configuration possibilities as well as a result set that contains information about the number of iterations. While Weka is written in Java, R and Julia are both written in their own high-level syntax. Even though, critical code parts are implemented in C, C++ and Fortran for better performance. Without the overhead of an entire database system behind, it will be interesting to compare HyPer with Weka, R and Julia.

We are excluding the big data solutions, middleware tools and analytical databases presented in Chapter 5, since these tools are often only applicable for specific use cases, e.g. tremendously large volumes of data, or require specific environments or proprietary software. Instead, we are comparing the HyPer k-Means solution with well-known, open-source clustering tools used by data scientists everyday.

The usability and user experience among R and Julia is very similar: Both can be run as interactive program or as script in their functional language, and both can import the test data sets in csv format. HyPer can be run as interactive console program or as script, too, and provides efficient csv loading [Müh+13]. Weka provides with the Weka Explorer a GUI, and can be run from the command line, too, or directly in Java programs. Internally, Weka uses the arff format for its algorithms. Using the Weka GUI to convert the original data (in our case in csv format) into the arff format feels fairly natural, as it happens in a preprocessing step when loading the data. However, when using the command line or Java programs to read a csv file, the file has to be preprocessed to the arff format first by an additional command. This takes time and is not as convenient compared to the other tools.

However, the main reason for not including Weka in the next sections is the disappointing running time compared to Julia, R and HyPer, as depicted in Table 7.2. The table shows the median, the 90th and the 95th percentile of the running time per iteration of 100 Weka clusterings for $k = 3, 10$ and 20. Compared to the other tools, which we will present in later sections in detail, Weka performs poorly: On the network data set, Weka is still slower by a factor of 7 compared to Julia's implementation, which is

already the slowest tool on that data set. On the high dimensional data set, the HyPer C++ implementation shows the worst results regarding Julia, R and HyPer, however, Weka is slower by a factor of around 6. As last test, the medium size data set was run against Weka, and it is slower by a factor of 10 for $k = 3$, 8 for $k = 10$ and 7 for $k = 20$ compared to Julia, the slowest one regarding this data set. Because of this tremendous differences in running time, we stopped the experiments with Weka and decided not to include Weka in the following sections and only compare the performance of HyPer operator to R and Julia.

Table 7.2.: Weka Results - Time per Iteration in seconds.

		Network 3D			High Dimensional			Medium Size	
k	3	10	20	3	10	20	3	10	20
50	1.39	1.54	1.93	1.08	1.39	1.90	54.01	58.25	71.11
90	1.48	1.60	2.02	1.16	1.44	1.98	59.16	62.54	87.16
95	1.49	1.61	2.06	1.19	1.46	2.01	59.79	64.31	88.54

percentile

7.3. Serial Implementation

In this section we compare the C++ driven and the LLVM driven implementation with each other, both described in detail in Section 6.4. The experiments differentiate between the compilation and the execution time. The compilation time is the time for generating LLVM code, while the execution time is the time for executing the k-Means algorithm after the LLVM code was generated. Since the compilation time can be seen as a one-time setup time, we expect the execution time to be much larger.

For each data set, the two HyPer implementations are tested for $k = 3, 10$ and 20. For each *k* the algorithm is executed 100 times with a maximum number of iterations of 10. As result, the median time per iteration in seconds is stated. For better comparability the compilation time is also relative to the number of iterations, even though it stays constant as the number of iterations grows.

Table 7.3 shows the result of the two serial implementations for the network data set. As we observe the compilation time is independent of the cluster number *k*. Also the compilation time does not differ among the C++ implementation and the LLVM version. In contrast, the execution time increases as the cluster number *k* increases. This is true for both implementations.

For better visualization, Figure 7.1 depicts the same result as stacked bar charts. Again we observe that the compilation time is low and constant, while the execution

Table 7.3.: 3D Network - Time per Iteration.

k	HyPer C++ compilation[s]	execution[s]	HyPer LLVM compilation[s]	execution[s]
3	0.0050	0.0680	0.0046	0.0171
10	0.0044	0.0947	0.0046	0.0502
20	0.0045	0.1391	0.0046	0.0901

time differs among the implementations. For $k = 3$, the execution time of the C++ version is almost four times the execution time of the LLVM version. For $k = 10$ it is two times the execution time and for $k = 20$ it is 1.5 times, respectively. On the other hand that means that for a larger cluster number k, the LLVM execution time grows faster than the execution time of the C++ version. Actually, from $k = 3$ to $k = 10$, the LLVM version grows by a factor of 2.9, while the C++ version grows by 1.4. From $k = 3$ to $k = 20$ the growth is even more significant, from factor 2 for the C++ version to 5.3 for the LLVM version.
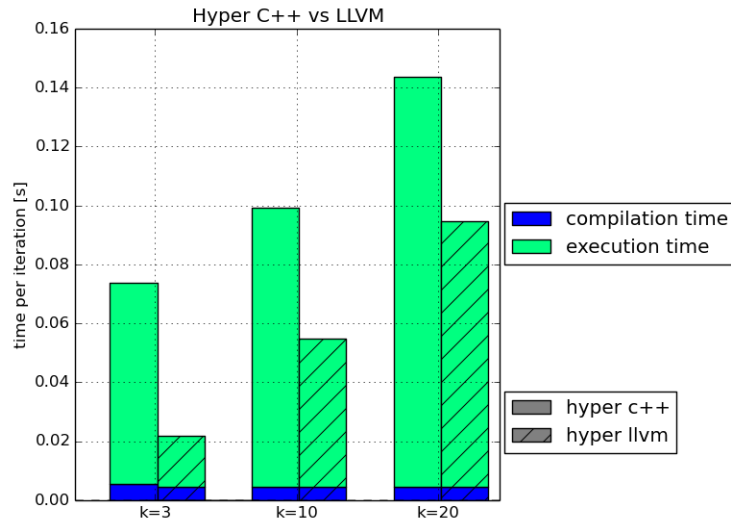

Figure 7.1.: 3D Network - Time per Iteration.

This gives us the first interesting results. Although the C++ version is implementing k-Means using LLVM only for generated functions, the compilation time does not differ compared to the LLVM version, where not only the functions but also the entire algorithm is written in LLVM. Nevertheless, the functions for computing the distance

and updating the centers are generated in LLVM for both implementations which is an explanation for the similar compilation times: In the C++ version this code is generated as functions callable from the `runtime system`, while the LLVM version embeds the code directly into the LLVM program structure. Therefore, the difference between the two seems to be insignificant.

Regarding the execution time, the LLVM version is much faster than the C++ version. One reason is that the C++ implementation has many function calls between the `compile time` and the `runtime system`. For the LLVM system these calls are not necessary, therefore the data can remain in the CPU registers. The second advantage is that the entire algorithm is compiled in LLVM code resulting in very efficient code, optimized on a lower level than even possible with C++ code.

Table 7.4.: High Dimensional - Time per Iteration.

k	HyPer C++ compilation[s]	execution[s]	HyPer LLVM compilation[s]	execution[s]
3	0.1278	0.0522	0.0933	0.0327
10	0.1287	0.1171	0.0933	0.0706
20	0.1299	0.2070	0.0933	0.1252

Table 7.4 shows the result of the same experiment with the high dimensional data set. In contrary to the network data set the compilation time is slightly different between the C++ and the LLVM implementation. Furthermore for both versions the compilation time is larger than the execution time for $k = 3$ and almost equal for $k = 10$. Only for $k = 20$ the execution time is larger than the compilation time, as depicted in Figure 7.2.

Again, the compilation time stays constant for different values of k. The increase of compilation time is induced by the high dimensionality: For each dimension additional code has to be generated. Therefore the code for a data set with four dimension is faster to compile than for a data set with 50 dimensions. The difference in the running time is similar to the network data set even though the increase in performance between the two implementations is not as significant. For $k = 3$, LLVM is faster by a factor 1.6 and for $k = 10$ and $k = 20$ by 1.7. This time the increase in execution time is strongly correlated. From $k = 3$ to $k = 10$, the LLVM version and the C++ version grow both by a factor of 2.2. From $k = 3$ to $k = 20$, the LLVM version grows by a factor of 4, while the C++ version grows by 3.8.

Interestingly, the execution time for the network data set is lower than for the high dimensional data set, even though the network set has a size of 0.019 GB, while the high dimensional data set has only 0.009 GB. Even if we omit the compilation time which stays constant for a growing number of iterations and only look at the execution
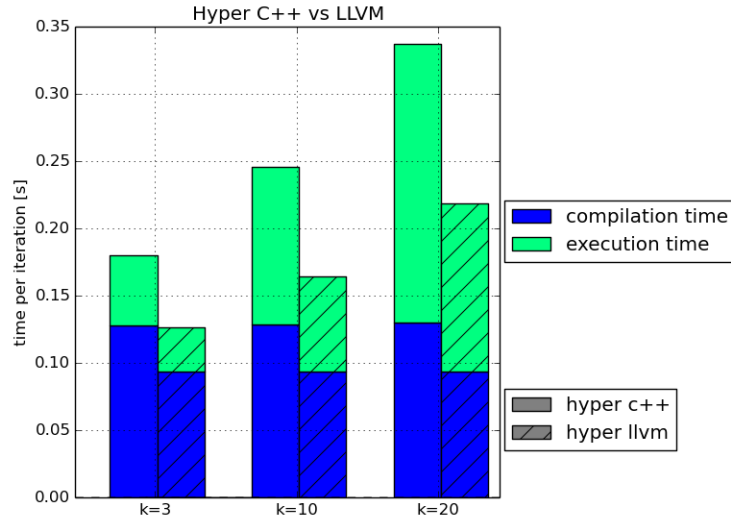
Figure 7.2.: High Dimensional - Time per Iteration.

time the network data set is only for the C++ implementation and $k = 3$ slower than the high dimensional data set. Apparently, our implementation performs much better for low dimensional data sets, where not that much code has to be generated. As already mentioned, for each additional dimension additional code has to be generated and executed. For small data sets this compilation time is already a limiting factor, in addition to the extended execution time.

Table 7.5 and Table 7.6 show the same experiment for the medium size and the medium size high dimensional data set. Both consist of 15 million instances, the first of four dimensions and the second of 50 dimensions. The compilation time of the medium size data set is similar to the network data set since both consist of four dimensions. The same is true for the medium size high dimensional data set and the previously used high dimensional data set: Both consist of 50 dimensions and the compilation time is quite similar. This shows us that the compilation time is independent of the data size and only affected by the number of dimensions. However, as the number of instances grows, the compilation time is not a significant factor for the overall running time anymore, as Figure 7.3 shows: The compilation time is not even visible plotting the data as stacked bar charts.

Again, the LLVM version shows a much better performance compared to the C++ version. For the medium size data set the overall running time is faster by a factor of 2.5, 1.6 and 1.4 for $k = 3, 10$ and 20, and for the medium size high dimensional data set by a factor of 1.6, 1.5 and 1.5, respectively.

Obviously the execution time for the high dimensional data set is much larger

Table 7.5.: Medium Size - Time per Iteration.

k	HyPer C++		HyPer LLVM	
	compilation[s]	execution[s]	compilation[s]	execution[s]
3	0.0041	2.3779	0.0047	0.9191
10	0.0041	3.2856	0.0047	2.1001
20	0.0041	4.6765	0.0047	3.4518

Table 7.6.: Medium Size High Dimensional - Time per Iteration.

k	HyPer C++		HyPer LLVM	
	compilation[s]	execution[s]	compilation[s]	execution[s]
3	0.1127	16.2787	0.0935	10.3302
10	0.1127	34.0051	0.0935	22.0209
20	0.1126	59.2904	0.0935	38.7046

compared to the low dimensional data set: This time the high dimensional data set has the same number of instances but 50 dimensions. For the C++ implementation the execution time is increased by a factor of 6.8 for $k = 3$, 10.3 for $k = 10$ and 12.6 for $k = 20$. For the LLVM implementation it is 11.2, 10.5, 11.2, respectively. Since the high dimensional data (2.79 GB) set is larger than the medium size data set (0.22 GB) by a factor of 12.7, the values are strongly correlated to the growth of the data set.

Finally, Table 7.7 shows the same experiment for the large size data set. The data set consists of ten dimensions and therefore the compilation time increases compared to the data sets with four dimensions by a factor of around two. However, since the data set has a size of 5.59 GB we can omit the compilation time since it does not affect the overall running time. Comparing the two versions regarding the overall running time, we observe that the LLVM version provides again much better results and is faster by a factor of 2.0, 1.5 and 1.4 for $k = 3, 10$ and 20 compared to the C++ implementation. This results in a difference of around 20 seconds per iteration between the two implementations.

In conclusion, we observed that the compilation time is not affected by the number of instances but the number of dimensions, where high dimensional data sets slow down the compilation phase. Interestingly, there is no significant difference regarding the compilation time between the two HyPer implementations.
Usually, the execution time outnumbers the compilation time by several factors. The only exception is the small, high dimensional data set as shown in the experiment.
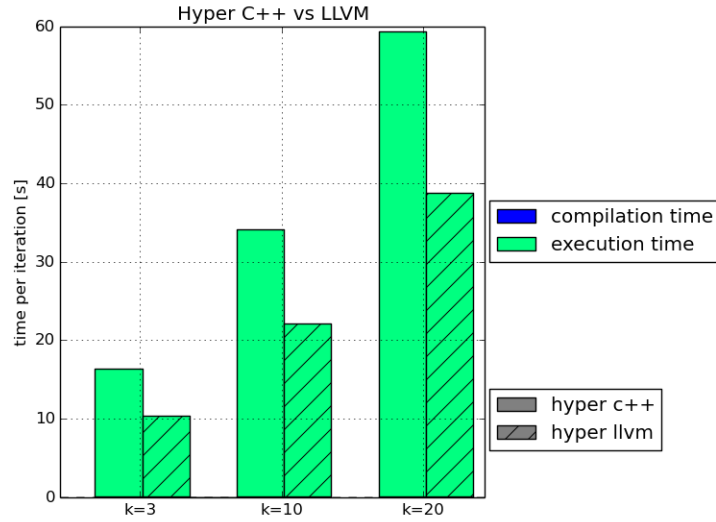
Figure 7.3.: Medium Size High Dimensional - Time per Iteration.

Table 7.7.: Large Size - Time per Iteration.

k	HyPer C++ compilation[s]	execution[s]	HyPer LLVM compilation[s]	execution[s]
3	0.0088	37.2804	0.0097	18.3734
10	0.0088	62.0034	0.0097	40.4783
20	0.0191	92.5868	0.0097	67.2364

Here, the compilation time is slower than the execution time for small *k*. Furthermore the execution time grows by the number of instances and is much faster for the LLVM implementation. With equality regarding the compilation and a performance increase regarding the execution time the LLVM version was the best choice for all performed experiments and the implementation aspects should be used as building block for future HyPer operators for data mining.

7.4. Performance Comparison with Competitor Systems

In this section we compare our serial HyPer implementations with R and Julia's implementation of the k-Means algorithm. As algorithm, the Lloyd algorithm with a maximum iteration number of ten is chosen. Apart from the large data set, the algorithm is executed 100 times for $k = 3, 10$ and 20, respectively. For the large data set,

the algorithm is executed ten times. The result is then presented as the execution time per iteration. For each algorithm the median, the 90th percentile and the 95th percentile are stated for each *k*.

Table 7.8.: 3D Network - Time per Iteration in seconds.

		Julia			R			HyPer C++			HyPer LLVM	
k	3	10	20	3	10	20	3	10	20	3	10	20
50	0.21	0.22	0.29	0.03	0.04	0.08	0.08	0.10	0.14	0.02	0.06	0.10
90	0.27	0.30	0.32	0.06	0.06	0.10	0.09	0.12	0.20	0.03	0.06	0.10
95	0.31	0.35	0.35	0.08	0.07	0.11	0.10	0.13	0.22	0.03	0.06	0.10

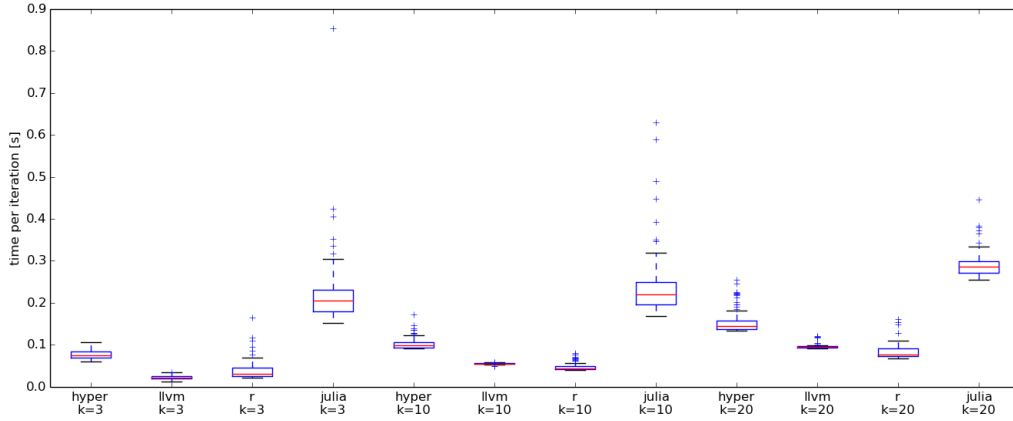(*percentile* labels the three rows 50, 90, 95)

Figure 7.4.: 3D Network - Time per Iteration.

First, we test against the real-world network data set. The result is presented in Table 7.8. As we already know, the HyPer LLVM implementation outnumbers the C++ version. For the median, the LLVM implementation is 3.4, 1.8 and 1.5 times faster for $k = 3, 10$ and 20. Julia performs the slowest, our LLVM implementation is 9.5, 4.0 and 3.1 times faster, respectively. Only the R implementation can compete with the HyPer LLVM operator and is with 0.7, 0.8 and 0.8 for $k = 3, 10$ and 20 even a bit faster. However, all tested programs differ in a few hundred milliseconds therefore the differences are not very significant.

Figure 7.4 depicts the result as a boxplot. We observe that the C++ version, the LLVM version and R are closely together, while Julia lays behind. Also regarding the variance of the 100 runs, Julia shows the strongest variations and many outliers.

Table 7.9.: High Dimensional - Time per Iteration in seconds.

		Julia			R			HyPer C++			HyPer LLVM	
k	3	10	20	3	10	20	3	10	20	3	10	20
percentile 50	0.04	0.05	0.07	0.03	0.05	0.08	0.18	0.24	0.35	0.13	0.16	0.22
90	0.04	0.05	0.07	0.04	0.06	0.10	0.21	0.34	0.40	0.13	0.16	0.22
95	0.04	0.05	0.07	0.05	0.07	0.11	0.22	0.35	0.44	0.13	0.17	0.22
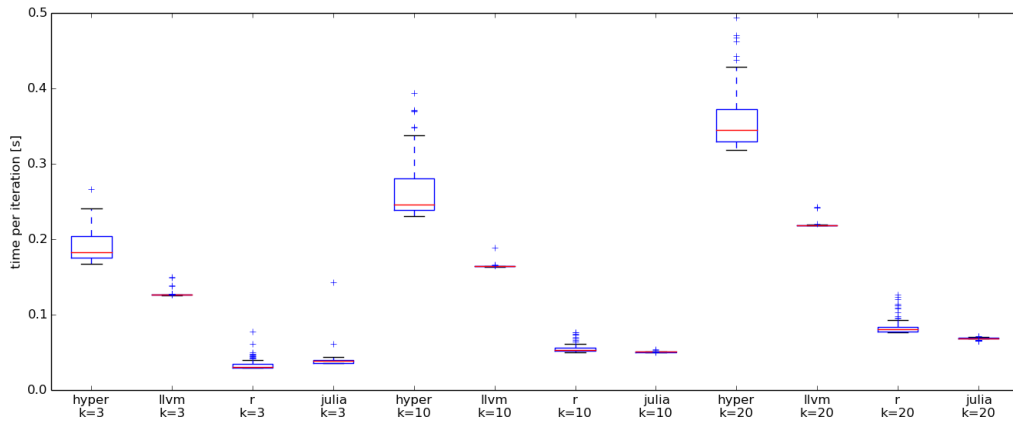
Figure 7.5.: High Dimensional - Time per Iteration.

Performing the same experiment on the high dimensional data set we see a different result, as depicted in Table 7.9. As already shown, the HyPer C++ and LLVM versions are slower compared to the network data set even though the data set is smaller by size. R takes almost the same time for both data sets. Interestingly, Julia's k-Means algorithm is now as fast as the R implementation, for $k = 10$ and 20 it is even faster. This is also depicted by the boxplot in Figure 7.5. Also the variance changed: The HyPer C++ version shows the highest variance with many outliers, while Julia's variance is very small this time.

This result gives us interesting knowledge about the different implementations. Both HyPer operators perform poorly when dimensions increase. An increase in dimensions and a decrease in instances does not affect R significantly. On the other hand, Julia shows much better results as dimensions increase and seems to be optimized for high dimensions: It is 5.4, 4.4 and 4.2 times faster for $k = 3, 10$ and 20 on the high dimensional set compared to the network data set, and the variance is much lower. A

conversation with Prof. Timothy E. Holy, Ph.D. substantiated this hypothesis[2].

Table 7.10.: Medium Size - Time per Iteration in seconds.

	Julia			R			HyPer C++			HyPer LLVM		
k	3	10	20	3	10	20	3	10	20	3	10	20
50	5.42	7.61	10.62	0.77	1.61	2.50	2.38	3.29	4.68	0.92	2.10	3.46
90	5.43	7.65	10.71	0.79	1.63	2.55	2.57	3.44	4.77	0.94	2.14	3.51
95	5.44	7.70	10.91	0.80	1.63	2.56	2.64	3.45	4.84	0.94	2.14	3.52

percentile

Table 7.11.: Medium Size HD - Time per Iteration in seconds.

	Julia			R			HyPer C++			HyPer LLVM		
k	3	10	20	3	10	20	3	10	20	3	10	20
50	11.46	15.69	21.67	9.43	15.48	23.77	16.39	34.12	59.40	10.42	22.11	38.80
90	11.51	15.81	21.69	9.50	15.50	23.79	16.77	36.11	59.53	10.44	22.12	38.84
95	11.65	15.83	21.70	9.64	15.58	23.79	17.29	42.15	59.57	10.45	22.13	38.84

percentile

We perform the same experiment on the medium size and the medium size high dimensional data set. Table 7.10 shows the result of the medium size data set. Again, the LLVM version is faster than the C++ version by a factor of 2.5, 1.6 and 1.4 for $k = 3, 10$ and 20. Since the dataset has only four dimensions Julia performs poorly and is by a factor of 5.9, 3.6 and 3.1 slower compared to the LLVM version. As for the network data set R demonstrates the best performance and is faster than LLVM by a factor of 0.8, 0.8 and 0.7 for $k = 3, 10$ and 20. Again we depict the result as a boxplot in Figure 7.6, where we observe a small variance for all used tools; only Julia has a few outliers for $k = 20$.

In comparison Table 7.11 and Figure 7.7 depict the result for the medium size high dimensional data set. As before, the HyPer implementation does not perform ideally on high dimensions. For $k = 3$, the LLVM version is slower than R by a factor of 0.9 but faster than Julia by a factor of 1.1. However, as k grows, R and Julia show slightly better results. Both show a speed up by a factor of 0.7 for $k = 10$, and 0.6 for $k = 20$, compared to the LLVM version.

Obviously the high dimensional data set takes more time per iteration having the same number of instances with a lot more dimensions. However, Julia shows a slow down by a factor of only 2 for all k's, although the dimensions increase from 4 to 50; R shows

---
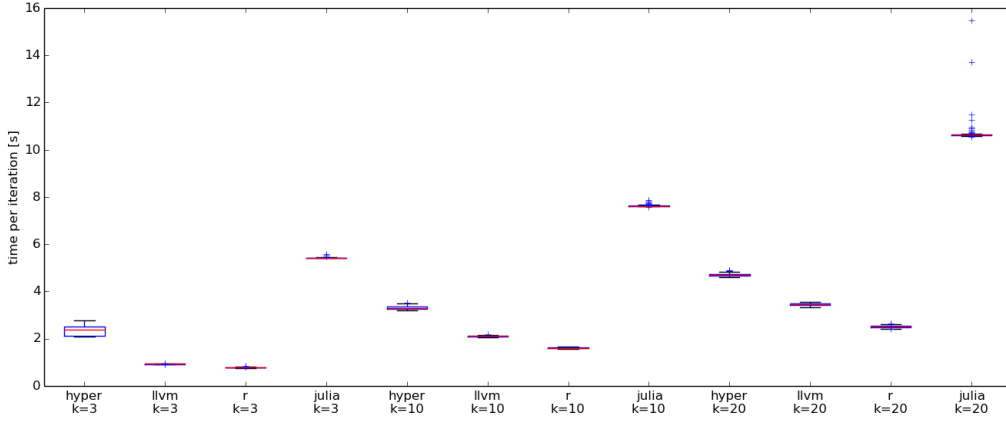
[2]Julia User Group: `http://bit.ly/1anXGMF`, 25.01.2015.

Figure 7.6.: Medium Size - Time per Iteration.

a slow down by a factor of 12 for $k = 3$, and 10 for $k = 10$ and 20, the C++ version by a factor of 7, 10, 13 for $k = 3, 10$ and 20, and the LLVM version by a factor of 11 for all k's. Julia performs again the best for high dimensional data. R and HyPer's LLVM operator behave very similar this time. An explanation is that the compilation time is not a limiting factor anymore for the LLVM implementation: For the high dimensional data set with 50,000 instances the compilation time was higher than the execution time. This time the data set consists of 15 million tuples therefore the poor compilation time does not affect the overall performance anymore.

As final test we perform the same experiment on the large data set containing 150 million instances, as shown in Table 7.12. This time the HyPer LLVM implementation outnumbers all the other tools, the C++ version by a factor of 2.0, 1.5 and 1.4 for $k = 3, 10$ and 20, Julia by a factor of 3.6, 2.2, 2.0, respectively, and even R by a factor of 1.6, 1.2 and 1.1, respectively. For growing k, the R implementation comes very close to the HyPer LLVM version. This result is also depicted in the boxplot in Figure 7.8.

In conclusion, the experiments demonstrate that the HyPer k-Means operator can compete with state-of-the-art tools for clustering. Particularly the LLVM implementation demonstrates good performance and shows similar results as the R implementation of the k-Means algorithm. However, for high dimensional data sets both HyPer operators do not perform ideally and are outnumbered by Julia's implementation optimized for high dimensions. Here is potential for future optimization as we demonstrate in Section 7.5 with a parallel implementation. Yet, if we take into account that the k-Means algorithm runs on top of an entire database with all its overhead, these results are supporting further development and the implementation of other data mining algorithms as HyPer operators.
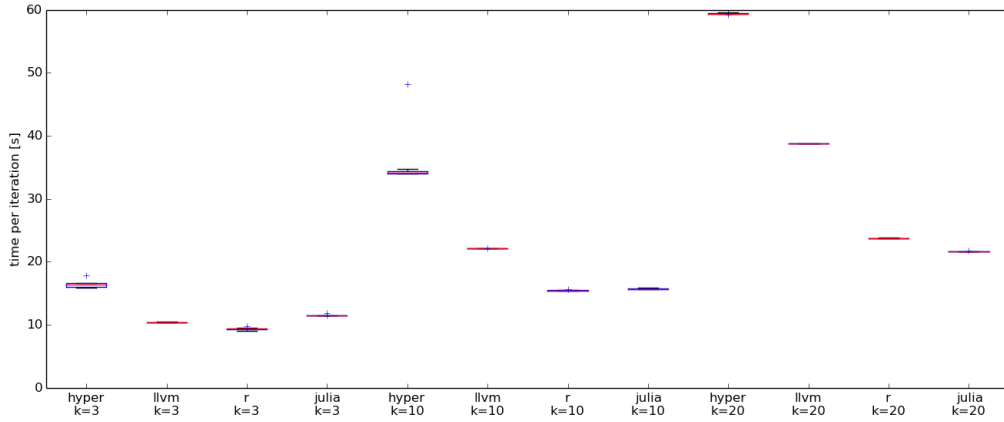
Figure 7.7.: Medium Size HD - Time per Iteration.

Table 7.12.: Large Size - Time per Iteration in seconds.

	Julia			R			HyPer C++			HyPer LLVM		
k	3	10	20	3	10	20	3	10	20	3	10	20
percentile 50	65.62	90.37	132.39	29.47	48.21	72.40	37.29	62.01	92.60	18.38	40.49	67.25
90	65.72	92.01	135.87	30.80	54.73	77.27	37.50	62.41	92.86	18.44	40.72	67.57
95	65.78	92.27	137.21	33.08	55.02	77.97	37.52	62.72	92.95	18.46	40.72	67.63

7.5. Parallel Implementation

So far we only looked at the serial implementations of the HyPer k-Means operator: In Section 7.3 we figured out that the LLVM implementation outperforms the C++ version on all used data sets. In Section 7.4 we compared the performance against Julia and R and showed that HyPer is able to compete with state-of-the-art technologies for data mining.

However, as data sets grow serial execution takes more and more time, making real-time data mining almost impossible. Furthermore, we do not exploit the possibilities of modern database hardware and the advantages of multi-threaded execution of algorithms. To address these challenges we presented a parallel implementation of the HyPer k-Means operator in Section 6.5. In this section we compare this parallel implementation with the LLVM version and the R implementation. Both tools demonstrate the best results on the medium size, the medium size high dimensional and the large data set. Julia and the C++ driven implementation are excluded in this section since their performance is either similar or slower to R and the LLVM implementation on the

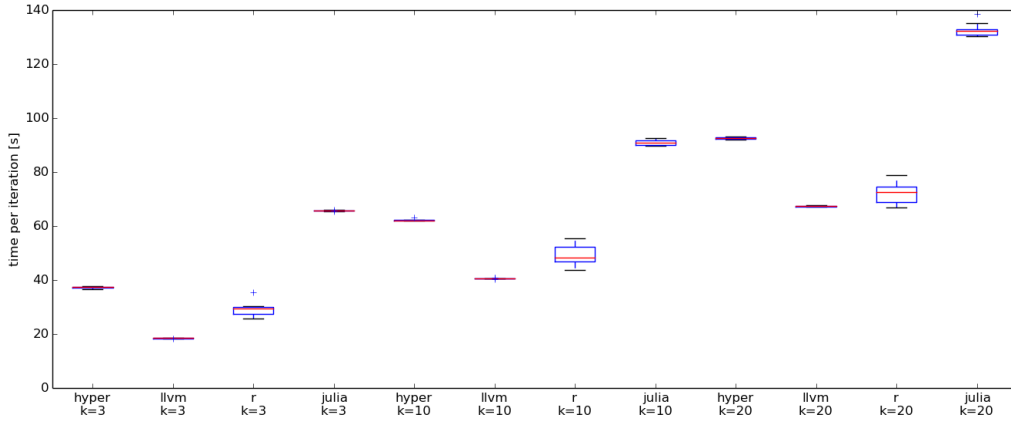Figure 7.8.: Large Size - Time per Iteration.

used data sets.

Table 7.13.: Medium Size - Time per Iteration in seconds.

	k	R 3	10	20	HyPer LLVM 3	10	20	HyPer Parallel 3	10	20
percentile	50	0.77	1.61	2.50	0.92	2.10	3.46	1.35	1.52	1.75
	90	0.79	1.63	2.55	0.94	2.14	3.51	1.38	1.53	1.84
	95	0.80	1.63	2.56	0.94	2.14	3.52	1.40	1.53	1.85

Table 7.13 shows the result as the time per iteration for $k = 3, 10$ and 20. As before, the median, the 90th and 95th percentile are presented. For $k = 3$, we surprisingly observe that both the R and the LLVM version are faster than the parallel version. However, as k grows, the parallel version outperforms R and the LLVM version. This circumstance is also shown in the bar chart in Figure 7.9.

The reason is that for the parallel execution the running time is almost independent of k because the computation is performed on different, independent threads that are only slightly affected by an increase of k. However, as k grows up to hundreds while the number of cores stays constant, the performance might be affected also in the parallel execution mode. For a k between 3 and 20 this does not happen and therefore the parallel version outperforms R and LLVM. For R, the time per iteration grows by a factor of 2.1 from $k = 3$ to $k = 10$, and by a factor of 1.6 from $k = 10$ to $k = 20$. For LLVM we observe an increase by a factor of 2.3 and 1.6, respectively. In contrast, the parallel implementation grows only by factor of 1.1 and 1.2, respectively. In other

words, R and the LLVM version grow by seconds as *k* grows, while the parallel version grows insignificantly by around 200 milliseconds.
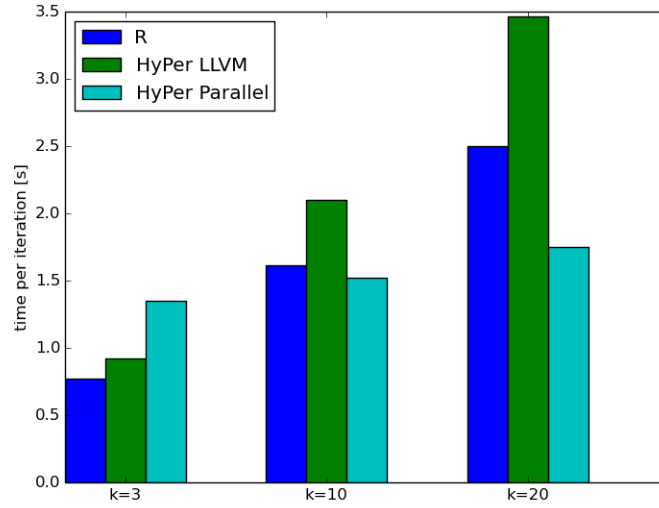

Figure 7.9.: Medium Size - Time per Iteration in seconds.

Table 7.14 shows the same experiment on the medium high dimensional data set. For this data set the parallel version outperforms the other two implementations for all *k*'s, even for $k = 3$. Figure 7.10 depicts that the parallel version is again almost independent of *k*, in contrast to R and LLVM. R grows by a factor of 1.6 from $k = 3$ to $k = 10$, and 1.5 from $k = 10$ to $k = 20$; for LLVM we observe an increase by a factor of 2.1 and 1.8, respectively, and for the parallel version a factor of 1.4 and 1.4, respectively. Even though the parallel version is not as independent by the cluster number *k* as for the low dimensional data set, the increase in time is much lower. Additionally, a performance gain is achieved by the ability to handle high dimensional data better: From four dimensions to 50 dimensions, the parallel execution time is only affected by a factor of 4.3, 5.3 and 6.5 for $k = 3, 10$ and 20. R and LLVM are affected by a much larger factor: 12.3, 9.6 and 9.5 for R, 11.3, 10.5 and 11.2 for the LLVM implementation. Again, the performance benefits from the parallel execution: Computing the distance from each data point to each center point is computationally expensive, even more when the instances consist of many dimensions. Then, a computation on many, independent threads speeds up the performance tremendously, as shown in this experiment.

Finally, Table 7.15 shows the result of the same experiment on the large size data set. Again, the parallel version is the fastest. For $k = 3$, the parallel version is faster than R by a factor of 1.8 and for LLVM by a factor of 1.1. For $k = 10$ the factor is even larger, 2.4 for R and 2.0 for LLVM, and 3.0 and 2.8 for $k = 20$, respectively. Since *k* affects the

Table 7.14.: Medium Size HD - Time per Iteration in seconds.

		R			HyPer LLVM			HyPer Parallel	
k	3	10	20	3	10	20	3	10	20
50	9.43	15.48	23.77	10.42	22.11	38.80	5.84	8.04	11.30
90	9.50	15.50	23.79	10.44	22.12	38.84	5.88	8.26	11.35
95	9.64	15.58	23.79	10.45	22.13	38.84	5.88	8.29	11.37

(percentile)
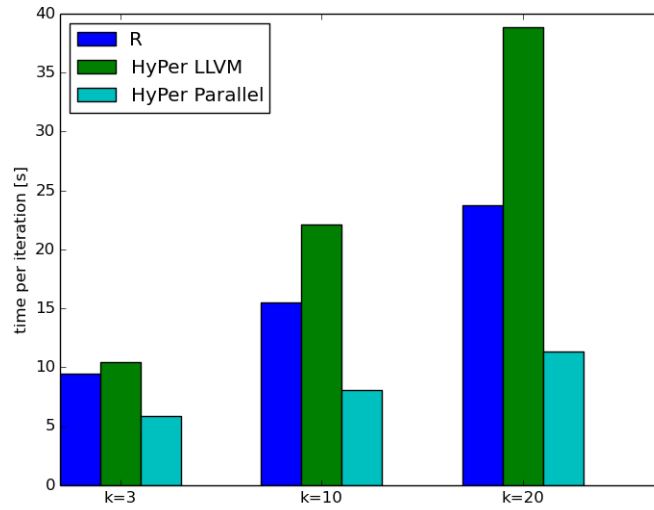
Figure 7.10.: Medium Size HD - Time per Iteration.

performance of the parallel version only slightly again, the factor is increasing as *k* increases.

To conclude, the parallel version performs much better than the LLVM and R implementation. Except for the medium size data set and $k = 3$ it demonstrates always better results. However, since we are using only 16 cores even a higher speed up would be possible, in particular for larger data sets, more dimensions and a growing *k*. Furthermore, we are using the serial C++ version of the HyPer k-Means operator as foundation for our parallel implementation. As seen in the previous sections, this implementation performs poorly compared to the LLVM implementation: As in the serial implementation we have many function calls between the `runtime` and the `compile time system` also in the parallel version. Additionally, we are using high-level C++ constructs over the more efficient LLVM code generation. Therefore, in future

implementations the LLVM version should be the building block of the parallel version which might lead to crucial performance improvements as observed already between the two serial implementations. And finally, only the computation of the distances is parallelized so far, while the recalculation of the cluster centers is still performed in serial. This is another potential of further improvement. Therefore, we can even expect better results when we enhance the addressed shortcomings in future versions.

Table 7.15.: Large Size - Time per Iteration in seconds.

		R			HyPer LLVM			HyPer Parallel	
k	3	10	20	3	10	20	3	10	20
50	29.47	48.21	72.40	18.38	40.49	67.25	16.71	19.79	24.03
90	30.80	54.73	77.27	18.44	40.72	67.57	16.94	20.13	24.33
95	33.08	55.02	77.97	18.46	40.72	67.63	17.01	20.17	24.36

percentile
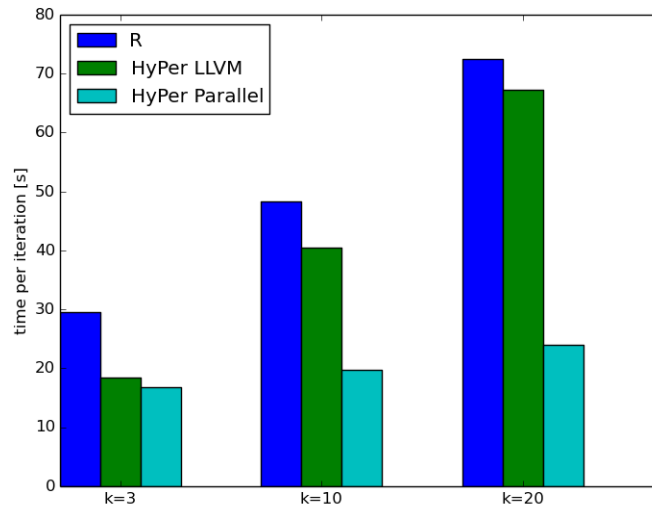
Figure 7.11.: Large Size - Time per Iteration.

8. Conclusions and Future Work

In this thesis we have presented a proof-of-concept implementation of the k-Means clustering algorithm as HyPer operator. This chapter provides conclusions of the presented work and addresses open challenges.

8.1. Conclusions

With main memory databases we are able to relax traditional database constraints and benefit from unprecedented performance gains. The main memory system HyPer can execute OLTP and OLAP queries on one system without the need of an additional data warehouse. Such a system provides the opportunity to execute data mining algorithms directly on the database, and to perform real-time data analysis.

In this thesis, we presented several approaches to implement the k-Means clustering algorithm as a proof-of-concept algorithm for data mining in HyPer: A C++ driven approach, an LLVM driven approach and a parallel implementation have been presented. The implementation details can be used as building blocks for further data mining algorithms.

In extensive experiments we observed that the LLVM version is faster than the C++ version on all used data sets, and that our operator is able to compete with existing data mining solutions such as R, Julia and Weka. In fact, the LLVM approach and the R implementation are usually providing the best results for serial execution of the k-Means algorithm. Only for high dimensional data sets, the HyPer operator presents slightly inferior results. However, the HyPer k-Means operator also provides the opportunity for parallel execution: As data sets grow, the parallel implementation is faster than all the other tools, also on high dimensional data sets.

In conclusion, we showed that the k-Means algorithm implemented as HyPer operator can compete with and even outperform existing solutions, therefore data mining using the HyPer main memory database is strongly practicable.

8.2. Future Work

Many open challenges remain for data mining in HyPer. In this section we discuss future work and begin with general improvements of the k-Means algorithm. The presented k-Means operator implementation has three input parameters: The cluster number k, the maximum number of iterations and a verbose option to output statistics of the performed execution, such as the center coordinates and the cluster compactness. In future implementations we want to enhance the capabilities of k-Means and make our algorithm more flexible. One approach is to allow the user to choose the distance function. So far, the k-Means operator implements the euclidean distance, however, for some problems, other distance measurements are more appropriate, such as the manhattan, euclidean, minkowski or chebyshev distance, or the cosine similarity.

Applying normalization before performing the k-Means clustering is often necessary to get a reasonable good result. A normalization can be either implemented within the k-Means operator or even better as standalone operator so that other data mining algorithms can benefit from it too.

For large data sets the execution time of the k-Means algorithm can be quite high since it might take many iterations until convergence. Usually, k-Means terminates when none of the clusters is changing anymore. However, for large data sets this could take a while, in particular when a few data points are still moving from cluster to cluster while the majority of data points remains constant. Nevertheless, an additional iteration of the algorithm is initiated, affecting the performance of the algorithm. An advancement is an input parameter specifying a tolerable change of data points, as realized in Julia's k-Means implementation. If the number of changing data points is underneath this tolerance threshold, the algorithm terminates even though there is a chance to find a slightly better clustering.

Heterogeneous input data is another issue of the k-Means algorithm. Not all data sets consist only of numerical data. Often it is a mixture between numerical, binary and categorical attributes. An approach is to first convert the binary and categorical attributes to numerical attributes, and then provide an extended distance and cost function to keep the different dimensions comparable, as proposed in k-Means Mixed [AD07]. Another option is to assign a weight to each dimension after the conversion, to deal with the different characteristics of binary and categorical attributes. This is not only useful for mixed data sets, but also for homogeneous numerical data sets.

K-Means is usually implemented as the Lloyd [Llo82] algorithm, resulting in reasonable good results. However, the execution time can be improved implementing the Hartigan and Wong algorithm [HW79]. In contrast to the Lloyd algorithm it updates the center points at any time a data point is moved and makes time-saving choices for finding the closest cluster. Therefore, such an implementation might improve the performance of

the k-Means operator even more.

The result of the HyPer k-Means clustering is currently presented as a relational table with an additional column specifying the cluster identifier. Since this table is returned by the `consume` function, it can be used in further SQL statements. The statistical information however, such as the final center coordinates or the number of iterations, is only printed to the console. Therefore this information is lost and cannot be used in further analytical queries. An improvement is to generate additional relational tables storing the statistical information of the performed clustering. However, up to now HyPer is only capable of pushing tuples to the next `consume` operator and not of generating additional tables. Therefore, an extension of the existing operator model is suggested: Almost all data mining algorithms produce several intermediate and final results, thus the creation of additional tables is an important feature for computational data mining in HyPer.

As database system, HyPer makes already extensive use of index structures, e.g. for efficient lookup and sorting. Data mining algorithms can also benefit from index structures: A nearest-neighbor data structure could improve clustering high dimensional data and boost the performance even more.

One of the main observations of Chapter 7 is that the LLVM implementation is much faster than the C++ version on all performed experiments. Therefore, the center initialization using the k-Means++ algorithm could benefit from a generated LLVM implementation instead of the current C++ implementation with slow `getDistance` calls between the `compile time` and the `runtime system`. Moreover, not only the k-Means++ implementation can benefit from an LLVM implementation, but also the parallel execution. As we have shown, for small k and medium size data sets the parallel execution is still slower than the LLVM implementation. The main reason are the calls from the `runtime` to the `compile time system`, as the C++ driven approach is the building block of the parallel implementation. Therefore, an implementation in LLVM code could tremendously boost the performance of the algorithm.

Furthermore the parallelization itself can be improved: Only the process of computing the distances is parallelized so far. While this is already a big performance improvement the algorithm might benefit even more from a parallel execution of the initialization process and of the recalculation of the cluster centers.

Obviously more data mining algorithms have to be implemented in HyPer to enable a full data mining experience. Therefore, the existing k-Means algorithm can be used as a building block. A similar partitional clustering algorithm is the k-Medoids algorithm [KR90]. In contrary to k-Means it chooses data points as centers instead of the mean of the mass of all assigned data points. This makes the algorithm more robust against outliers. For implementation, the existing k-Means algorithm can be reused, only the way of computing center points after data point assignment has to be changed

according to the requirements of k-Medoids. As long term goal, we plan to include more clustering algorithms as well as algorithms for mining frequent itemsets and association rules, classification and outlier detection.

For finding a ready market in the data science community it will not be enough to provide HyPer with state-of-the-art data mining algorithms; it will also be necessary to provide a functional language for invoking the k-Means operator and other data mining algorithms, since not all data scientists might be fluent in the SQL language. Moreover, not only a functional language has to be provided, but also a way to visualize the results of data mining algorithms: The HyPer system and its functional language must be able to connect with a 2D plotting library to produce publication-ready, quality figures in a variety of forms. An example for a 2D plotting library based on a functional language is the Python library matplotlib [Hun07].

A. Data Set Generation Script

```python
import numpy as np
import argparse

parser = argparse.ArgumentParser(description='Generate a file for clustering')
parser.add_argument('rows', help="number of rows", type=int)
parser.add_argument('cols', help="number of columns", type=int)
parser.add_argument('output', help="output file name")

args = parser.parse_args()

data = np.random.randint(1, 99, (args.rows, args.cols))
data = data / 10.0

np.savetxt(args.output, data, fmt='%.1f', delimiter=",")
```

Acronyms

ACID Atomicity, Consistency, Isolation and Durability.

API Application programming interface.

arff Attribute-Relation File Format.

ASCII American Standard Code for Information Interchange.

CPU Central processing unit.

CRAN Comprehensive R Archive Network.

csv Comma-separated values.

DBMS Database Management System.

DBSCAN Density-Based Spatial Clustering of Applications with Noise.

ELKI Environment for Developing KDD-Applications Supported by Index-Structures.

ETL Extract, Transform, Load.

GB Gigabyte.

GUI Graphical User Interface.

HANA formerly High Performance Analytic Appliance.

HD High Dimensional.

HDFS Hadoop Distributed File System.

html Hypertext Markup Language.

HyPer Hybrid OLTP & OLAP High Performance.

ICDM International Conference on Data Mining.

IDE Integrated Development Environment.

IEEE Institute of Electrical and Electronics Engineers.

IT Information Technology.

JGR Java GUI for R.

JIT Just-in-time.

KDD Knowledge Discovery in Databases.

LLVM formerly Low Level Virtual Machine.

MB Megabyte.

NP-hard Non-deterministic Polynomial-time hard.

ODM Oracle Data Mining.

OLAP Online Analytical Processing.

OLTP Online Transaction Processing.

ORE Oracle R Enterprise.

PAL Predictive Analysis Library.

SQL Structured Query Language.

STING A Statistical Information Grid Approach to Spatial Data Mining.

SVM Support Vector Machine.

TPC Transaction Processing Performance Council.

TU Technische Universität.

TUM Technische Universität München.

UCI University of California, Irvine.

Weka Waikato Environment for Knowledge Analysis.

xml Extensible Markup Language.

List of Figures

List of Tables

Bibliography

[AD07]   A. Ahmad and L. Dey. "A k-mean clustering algorithm for mixed numeric and categorical data." In: *Data Knowl. Eng.* 63.2 (2007), pp. 503–527.

[Agr+98]   R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications." In: *SIGMOD Rec.* 27.2 (June 1998), pp. 94–105. ISSN: 0163-5808. DOI: 10.1145/276305.276314.

[AKZ08]   E. Achtert, H. Kriegel, and A. Zimek. "ELKI: A Software System for Evaluation of Subspace Clustering Algorithms." In: *Scientific and Statistical Database Management, 20th International Conference, SSDBM 2008, Hong Kong, China, July 9-11, 2008, Proceedings.* 2008, pp. 580–585. DOI: 10.1007/978-3-540-69497-7_41.

[Alo+09]   D. Aloise, A. Deshpande, P. Hansen, and P. Popat. "NP-hardness of Euclidean Sum-of-squares Clustering." In: *Mach. Learn.* 75.2 (May 2009), pp. 245–248. ISSN: 0885-6125. DOI: 10.1007/s10994-009-5103-0.

[Apa]   Apache Software Foundation. *Apache Mahout: Scalable machine-learning and data-mining library.*

[AS94]   R. Agrawal and R. Srikant. "Fast Algorithms for Mining Association Rules in Large Databases." In: *20th International Conference on Very Large Data Bases.* Morgan Kaufmann, Los Altos, CA, 1994, pp. 478–499.

[Aul+08]   S. Aulbach, T. Grust, D. Jacobs, A. Kemper, and J. Rittinger. "Multi-tenant Databases for Software As a Service: Schema-mapping Techniques." In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data.* SIGMOD '08. Vancouver, Canada: ACM, 2008, pp. 1195–1206. ISBN: 978-1-60558-102-6. DOI: 10.1145/1376616.1376736.

[Aul+09]   S. Aulbach, D. Jacobs, A. Kemper, and M. Seibold. "A Comparison of Flexible Schemas for Software As a Service." In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data.* SIGMOD '09. Providence, Rhode Island, USA: ACM, 2009, pp. 881–888. ISBN: 978-1-60558-551-2. DOI: 10.1145/1559845.1559941.

[AV07]   D. Arthur and S. Vassilvitskii. *k-means++: the advantages of carefull seeding*. 2007, pp. 1027–1035.

[Ber02]   P. Berkhin. *Survey of clustering data mining techniques*. Technical report. Accrue Software, San Jose, CA, 2002.

[Bez+12]   J. Bezanson, S. Karpinski, V. B. Shah, and A. Edelman. "Julia: A Fast Dynamic Language for Technical Computing." In: *CoRR* abs/1209.5145 (2012).

[BGV92]   B. E. Boser, I. M. Guyon, and V. N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers." In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: ACM, 1992, pp. 144–152. ISBN: 0-89791-497-X. DOI: 10.1145/130385. 130401.

[DG08]   J. Dean and S. Ghemawat. "MapReduce: Simplified Data Processing on Large Clusters." In: *Commun. ACM* 51.1 (Jan. 2008), pp. 107–113. ISSN: 0001-0782. DOI: 10.1145/1327452.1327492.

[Est+96]   M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In: *Second International Conference on Knowledge Discovery and Data Mining*. Ed. by E. Simoudis, J. Han, and U. M. Fayyad. AAAI Press, 1996, pp. 226–231.

[FA10]   A. Frank and A. Asuncion. *UCI machine learning repository*. 2010.

[Fär+12]   F. Färber, S. K. Cha, J. Primsch, C. Bornhövd, S. Sigg, and W. Lehner. "SAP HANA Database: Data Management for Modern Business Applications." In: *SIGMOD Rec.* 40.4 (Jan. 2012), pp. 45–51. ISSN: 0163-5808. DOI: 10.1145/ 2094114.2094126.

[FF48]   R. A. ; Y. Fisher and Frank. *[1938]*. Ed. by S. tables for biological, agri-cultural, and medical research (3rd ed.) London: Oliver & Boyd, 1948, pp. 26–27.

[Hal+09]   M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The WEKA Data Mining Software: An Update." In: *SIGKDD Explor. Newsl.* 11.1 (Nov. 2009), pp. 10–18. ISSN: 1931-0145. DOI: 10.1145/ 1656274.1656278.

[Han+04]   J. Han, J. Pei, Y. Yin, and R. Mao. "Mining Frequent Patterns Without Candidate Generation: A Frequent-Pattern Tree Approach." In: *Data Min. Knowl. Discov.* 8.1 (Jan. 2004), pp. 53–87. ISSN: 1384-5810. DOI: 10.1023/B: DAMI.0000005258.31418.83.

[Hel+12]   J. M. Hellerstein, C. Ré, F. Schoppmann, D. Z. Wang, E. Fratkin, A. Gorajek, K. S. Ng, C. Welton, X. Feng, K. Li, and A. Kumar. "The MADlib Analytics Library: Or MAD Skills, the SQL." In: *Proc. VLDB Endow.* 5.12 (Aug. 2012), pp. 1700–1711. ISSN: 2150-8097. DOI: 10.14778/2367502.2367510.

[HKP11]   J. Han, M. Kamber, and J. Pei. *Data Mining: Concepts and Techniques.* 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN: 0123814790, 9780123814791.

[Hun07]   J. D. Hunter. "Matplotlib: A 2D graphics environment." In: *Computing In Science & Engineering* 9.3 (2007), pp. 90–95.

[HW79]   J. A. Hartigan and M. A. Wong. "Algorithm AS 136: A k-means clustering algorithm." In: *In: Applied Statistics 28* 1 (1979), pp. 100–108.

[J+01]   E. Jones, T. Oliphant, P. Peterson, et al. *SciPy: Open source scientific tools for Python.* [Online; accessed 2015-02-28]. 2001–.

[KN11]   A. Kemper and T. Neumann. "HyPer: A hybrid OLTP amp;OLAP main memory database system based on virtual memory snapshots." In: *Data Engineering (ICDE), 2011 IEEE 27th International Conference on.* Apr. 2011, pp. 195–206. DOI: 10.1109/ICDE.2011.5767867.

[KR90]   L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis.* New York: John Wiley and Sons, 1990.

[KYJ13]   M. Kaul, B. Yang, and C. Jensen. "Building Accurate 3D Spatial Networks to Enable Next Generation Intelligent Transportation Systems." In: *Mobile Data Management (MDM), 2013 IEEE 14th International Conference on.* Vol. 1. June 2013, pp. 137–146. DOI: 10.1109/MDM.2013.24.

[LA04]   C. Lattner and V. S. Adve. *LLVM: A compilationframework for lifelong program analysis & transformation.* InACM International Symposium onCode Generation and Optimization (CGO), pages75-88, 2004.

[Llo82]   S. P. Lloyd. "Least squares quantization in PCM." In: *IEEE Transactions on Information Theory* 28.2 (1982), pp. 129–136. DOI: 10.1109/TIT.1982.1056489.

[LM10]   A. Lakshman and P. Malik. "Cassandra: A Decentralized Structured Storage System." In: *SIGOPS Oper. Syst. Rev.* 44.2 (Apr. 2010), pp. 35–40. ISSN: 0163-5980. DOI: 10.1145/1773912.1773922.

[Lor74]   R. A. Lorie. *XRM - an extended (n-ary) relationalmemory.* IBM Research Report, 1974.

[Mac67]   J. B. MacQueen. *Some Methods for classification and Analysis of Multivari-ate Observations*. In: Berkeley Symposium on Mathematical Statistics and Probability, 1967.

[MNV09]   M. Mahajan, P. Nimbhorkar, and K. Varadarajan. "The Planar k-Means Problem is NP-Hard." In: *Proceedings of the 3rd International Workshop on Algorithms and Computation*. WALCOM '09. Kolkata, India: Springer-Verlag, 2009, pp. 274–285. ISBN: 978-3-642-00201-4. DOI: `10.1007/978-3-642-00202-1_24`.

[Müh+13]   T. Mühlbauer, W. Rödiger, R. Seilbeck, A. Reiser, A. Kemper, and T. Neu-mann. "Instant Loading for Main Memory Databases." In: *Proc. VLDB Endow.* 6.14 (Sept. 2013), pp. 1702–1713. ISSN: 2150-8097. DOI: `10.14778/2556549.2556555`.

[Müh+14]   T. Mühlbauer, W. Rödiger, R. Seilbeck, A. Reiser, A. Kemper, and T. Neu-mann. "One DBMS for All: The Brawny Few and the Wimpy Crowd." In: *Proceedings of the 2014 ACM SIGMOD International Conference on Manage-ment of Data*. SIGMOD '14. Snowbird, Utah, USA: ACM, 2014, pp. 697–700. ISBN: 978-1-4503-2376-5. DOI: `10.1145/2588555.2594527`.

[Neu11]   T. Neumann. "Efficiently Compiling Efficient Query Plans for Modern Hardware." In: *Proc. VLDB Endow.* 4.9 (June 2011), pp. 539–550. ISSN: 2150-8097. DOI: `10.14778/2002938.2002940`.

[Neu14]   T. Neumann. "Viktor Leis: Compiling Database Queries into Machine Code." In: *IEEE Data Eng. Bull* 37.1 (2014).

[Ora15]   Oracle. *Oracle Advanced Analytics Option*. Mar. 2015. URL: `http://www.oracle.com/technetwork/database/options/advanced-analytics/overview/index.html`.

[Pag+99]   L. Page, S. Brin, R. Motwani, and T. Winograd. "The PageRank Citation Ranking: Bringing Order to the Web." In: 1999-66 (Nov. 1999). Previous number = SIDL-WP-1999-0120.

[PCY95]   J. S. Park, M.-S. Chen, and P. S. Yu. "An Effective Hash-based Algorithm for Mining Association Rules." In: *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*. SIGMOD '95. San Jose, California, USA: ACM, 1995, pp. 175–186. ISBN: 0-89791-731-6. DOI: `10.1145/223784.223813`.

[Qui93]   J. R. Quinlan. *C4.5: Programs for Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN: 1-55860-238-0.

[R D08]   R Development Core Team. *R: A Language and Environment for Statistical Computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria, 2008.

[Rap14]   RapidMiner. *RapidMiner Studio 6.2 and Server 2.2*. 2014.

[Ros58]   F. Rosenblatt. "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain." In: *Psychological Review* 65.6 (1958), pp. 386–408.

[SAP14]   SAP. *SAP HANA Predictive Analysis Library*. Nov. 2014.

[Shv+10a]   K. Shvachko, H. Kuang, S. Radia, and R. Chansler. "The Hadoop Distributed File System." In: *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. MSST '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–10. ISBN: 978-1-4244-7152-2. DOI: `10.1109/MSST.2010.5496972`.

[Shv+10b]   K. Shvachko, H. Kuang, S. Radia, and R. Chansler. "The Hadoop Distributed File System." In: *Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*. MSST '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 1–10. ISBN: 978-1-4244-7152-2. DOI: `10.1109/MSST.2010.5496972`.

[SON95]   A. Savasere, E. Omiecinski, and S. B. Navathe. "An Efficient Algorithm for Mining Association Rules in Large Databases." In: *Proceedings of the 21th International Conference on Very Large Data Bases*. VLDB '95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 432–444. ISBN: 1-55860-379-4.

[Thu+09]   A. Thusoo, J. S. Sarma, N. Jain, Z. Shao, P. Chakka, S. Anthony, H. Liu, P. Wyckoff, and R. Murthy. "Hive: A Warehousing Solution over a Mapreduce Framework." In: *Proc. VLDB Endow.* 2.2 (Aug. 2009), pp. 1626–1629. ISSN: 2150-8097. DOI: `10.14778/1687553.1687609`.

[Wu+08]   X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand, and D. Steinberg. "Top 10 algorithms in data mining." English. In: *Knowledge and Information Systems* 14.1 (2008), pp. 1–37. ISSN: 0219-1377. DOI: `10.1007/s10115-007-0114-2`.

[WYM97]   W. Wang, J. Yang, and R. R. Muntz. "STING: A Statistical Information Grid Approach to Spatial Data Mining." In: *Proceedings of the 23rd International Conference on Very Large Data Bases*. VLDB '97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 186–195. ISBN: 1-55860-470-7.

[Zah+10]   M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. "Spark: Cluster Computing with Working Sets." In: *Proceedings of the 2Nd USENIX Conference on Hot Topics in Cloud Computing*. HotCloud'10. Boston, MA: USENIX Association, 2010, pp. 10–10.

[ZMH09]   W. Zhao, H. Ma, and Q. He. "Parallel K-Means Clustering Based on MapReduce." English. In: *Cloud Computing*. Ed. by M. Jaatun, G. Zhao, and C. Rong. Vol. 5931. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2009, pp. 674–679. ISBN: 978-3-642-10664-4. DOI: 10.1007/978-3-642-10665-1_71.