The Journal of Systems and Software 134 (2017) 170–189

Contents lists available at ScienceDirect

The Journal of Systems and Software

journal homepage: www.elsevier.com/locate/jss

Task Scheduling in Big Data Platforms: A Systematic Literature Review

Mbarka Soualhia[a,*], Foutse Khomh[b], Sofiène Tahar[a]

[a] Concordia University, Canada
[b] Polytechnique Montréal, Montréal, Quebec, Canada

ARTICLE INFO

Article history:
Received 20 October 2016
Revised 18 July 2017
Accepted 1 September 2017
Available online 5 September 2017

Keywords:
Task Scheduling
Hadoop
Spark
Storm
Mesos
Systematic Literature Review

ABSTRACT

Context: Hadoop, Spark, Storm, and Mesos are very well known frameworks in both research and industrial communities that allow expressing and processing distributed computations on massive amounts of data. Multiple scheduling algorithms have been proposed to ensure that short interactive jobs, large batch jobs, and guaranteed-capacity production jobs running on these frameworks can deliver results quickly while maintaining a high throughput. However, only a few works have examined the effectiveness of these algorithms.

Objective: The Evidence-based Software Engineering (EBSE) paradigm and its core tool, *i.e.*, the Systematic Literature Review (SLR), have been introduced to the Software Engineering community in 2004 to help researchers systematically and objectively gather and aggregate research evidences about different topics. In this paper, we conduct a SLR of task scheduling algorithms that have been proposed for big data platforms.

Method: We analyse the design decisions of different scheduling models proposed in the literature for Hadoop, Spark, Storm, and Mesos over the period between 2005 and 2016. We provide a research taxonomy for succinct classification of these scheduling models. We also compare the algorithms in terms of performance, resources utilization, and failure recovery mechanisms.

Results: Our searches identifies 586 studies from journals, conferences and workshops having the highest quality in this field. This SLR reports about different types of scheduling models (dynamic, constrained, and adaptive) and the main motivations behind them (including data locality, workload balancing, resources utilization, and energy efficiency). A discussion of some open issues and future challenges pertaining to improving the current studies is provided.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

The processing and analysis of datasets in cloud environments has become an important and challenging problem, because of the exponential growth of data generated by social networks, research and healthcare platforms, just to name a few. Hadoop (Kurazumi et al., 2012), Spark (Zaharia et al., 2010b), Storm (Peng et al., 2015a), and Mesos (Hindman et al., 2011b) are examples of widely used frameworks for distributed storage and distributed processing of ultra large data-sets in the cloud. Many large organisations like Yahoo!, Google, IBM, Facebook, or Amazon have deployed these well-known big data frameworks (Jian et al., 2013a). Hadoop, Spark, Storm, and Mesos are multi-tasking frameworks that support a variety of different types of tasks processing. They have a pluggable architecture that permits the use of schedulers optimized for particular workloads and applications. The scheduling of tasks in these frameworks is of a paramount importance since it affects the computation time and resources utilisation. However, because of the dynamic nature of cloud environments, efficient task scheduling is very challenging. Multiple algorithms have been proposed to improve how tasks are submitted, packaged, scheduled and recovered (in case of failures) in these frameworks. Yet, only a few works have compared the proposed algorithms and investigated their impact on the performance of the aforementioned frameworks. To the best of our knowledge, there is no published literature that clearly articulates the problem of scheduling in big data frameworks and provides a research taxonomy for succinct classification of the existing scheduling techniques in Hadoop, Spark, Storm, and Mesos frameworks. Previous efforts (Patil and Deshmukh, 2012), (Rao and Reddy, 2012; Singh and Agrawal, 2015) that attempted to provide a comprehensive review of scheduling issues in big data platforms were limited to Hadoop only. Moreover, they did not include all papers that were published during the periods covered by their studies (*i.e.*, 2012

* Corresponding author.
E-mail addresses: soualhia@ece.concordia.ca (M. Soualhia), foutse.khomh@polymtl.ca (F. Khomh), tahar@ece.concordia.ca (S. Tahar).

http://dx.doi.org/10.1016/j.jss.2017.09.001
0164-1212/© 2017 Elsevier Inc. All rights reserved.

and 2015). Also, these three studies only propose general descriptions of Hadoop schedulers in terms of architecture and objectives (*e.g.*, learning, resources management) and do not discuss their limitations. Neither do they discuss future research directions to improve these existing task scheduling approaches.

In this paper, we follow the Evidence-based Software Engineering (EBSE) paradigm in order to conduct a Systematic Literature Review (SLR) (Kitchenham, 2004) of task scheduling techniques in Hadoop, Spark, Storm, and Mesos, with the aim to identify and classify the open challenges associated with task scheduling in these frameworks.

We discuss different approaches and models of task scheduling proposed for these four frameworks, that gained a lot of momentum in the last decade in both research and commercial communities. Also, we analyse the proposed design decision of each approach in terms of performance, resources utilization, failure recovery mechanisms, and energy efficiency. Our searches identified 586 journals, conferences and workshops papers published in top ranked software engineering venues between 2005 and 2016. We organize our SLR in three parts:

- ***Part 1:*** *Task Scheduling Issues in Big Data Platforms:*

First, we present the main issues related to task scheduling in Hadoop, Spark, Storm, and Mesos, and explain how these issues are addressed by researchers in the existing literature. We classify the issues into 6 main categories as follows: *resources management, data management* (including *data locality, replication* and *placement* issues), *fairness, workload balancing, fault-tolerance*, and *energy-efficiency*.

- ***Part 2:*** *Task Scheduling Solutions in Big Data Platforms:*

Second, we describe the different types of scheduling approaches available in the open literature and discuss their impact on the performance of the schedulers of the four frameworks. Overall, we observe that we can classify the scheduling approaches used in Hadoop, Spark, Storm, and Mesos into three main categories: *dynamic, constrained* and *adaptive* scheduling.

- ***Part 3:*** *Research Directions on Task Scheduling in Big Data Platforms:*

Third, we describe some of the future research directions that can be addressed in each category discussed previously in *part 1* and *part 2* of the SLR. From the limitations of previous work (discussed in *part 2*), we build a roadmap for future research to improve existing scheduling approaches.

The remainder of this paper is organized as follows: Section 2 briefly introduces Hadoop, Spark, Storm, and Mesos. Section 3 describes the methodology followed in this Systematic Literature Review. Sections 4, 5 and 6 discuss our study and the findings of this review, and position our work in the existing literature. Section 7 presents our conclusions, and outlines the main findings of this systematic review.

2. Background

Fig. 1 describes the relationships between MapReduce, Hadoop, Spark, Storm, and Mesos. Hadoop is a well-known processing platform that implements the MapReduce programming model. Spark is a novel in-memory computing framework that can be running on Hadoop. Storm is a distributed computation framework for real time applications. Spark and Storm can implement the MapReduce programming model, but with different features to handle their topologies and data models. These platforms can be typically deployed in a cluster, that can be managed by Mesos or YARN (Yet Another Resources Negotiator), which are cluster managers. In the sequel, we briefly describe MapReduce, Hadoop, Spark, Storm, and Mesos.

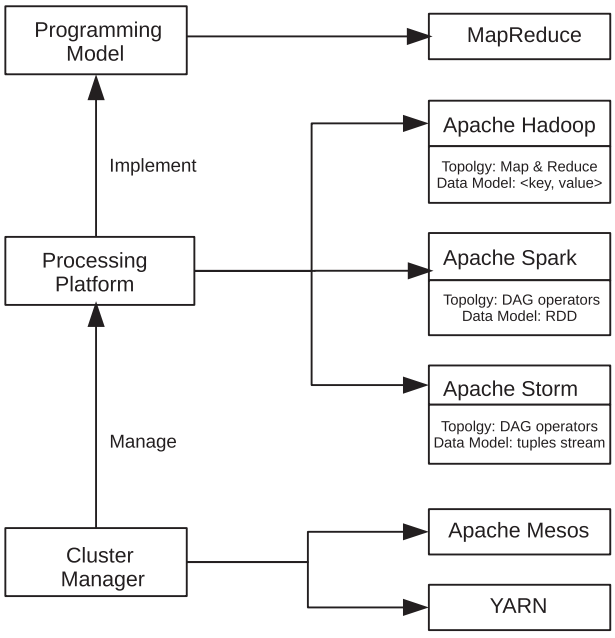

Fig. 1. An overview of relationships among MapReduce, Hadoop, Spark, Storm, and Mesos

2.1. Programming Model: MapReduce

MapReduce Lee et al. (2012) is a programming model for processing big amounts of data using a large number of computers (nodes). It subdivides the received users' requests into parallel jobs and executes them on processing nodes where data are located, instead of sending data to the nodes that execute the jobs. A MapReduce job is composed of *"map"* and *"reduce"* functions and the input data. The input data represents a set of distributed files that contain the data to be processed. The *map* and *reduce* functions are commonly used in functional programming languages like Lisp. The map function takes the input data and outputs a set of $<$key, value$>$ pairs. The reduce function takes the set of values for a given key as input and emits the output data for this key. A shuffling step is performed to transfer the map outputs to the corresponding reducers. The set of intermediate keys are sorted by Hadoop and given to the reducers. For each intermediate key, Hadoop passes the key and its corresponding sorted intermediate values to the reduce function. The reducers (*i.e.*, worker running a reduce function) use a hash function to collect the intermediate data obtained from the mappers (*i.e.*, worker running a map function) for the same key. Each reducer can execute a set of intermediate results belonging to the mappers at a time. The final output of the reduce function will be stored in a file in the distributed file system (Dean and Ghemawat, 2008). MapReduce follows a master-slave model. The master is known as *"JobTracker"*, which controls the execution of the "map" and "reduce" functions across the slave workers using *"TaskTrackers"*. The JobTracker and the TaskTrackers control the job execution to ensure that all functions are executed and have their input data as shown in Fig. 2.

2.2. Processing Platforms

In the sequel, we describe Hadoop, Spark, and Storm processing platforms and we briefly discuss task scheduling issues in these platforms.

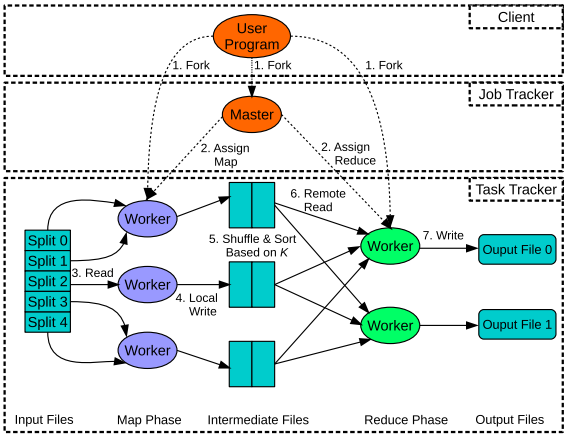

Fig. 2. An overview of job execution in MapReduce (Dean and Ghemawat, 2008).

2.2.1. Apache Hadoop

Hadoop (Had, 2017) has become the *de facto* standard for processing large data in today's cloud environments. It is a Java-based MapReduce implementation for large clusters that was proposed by Cutting and Cafarella in 2005 (Had, 2017). Hadoop is composed of two main components: the Hadoop Distributed File System (HDFS) and the MapReduce framework. The HDFS is responsible for storing and managing the input of the map function as well as the output of the reduce function. The Hadoop MapReduce framework follows a master-slave model (Dean and Ghemawat, 2008). The JobTracker running on the master is responsible for managing the job execution, progress and the status of the workers (slaves). Each worker in Hadoop is composed of a TaskTracker and a DataNode. The TaskTracker is responsible for processing the jobs using their corresponding input data located in the DataNode (Dean and Ghemawat, 2008). Hadoop allows the processing of large data-sets across a distributed cluster using a simple programming model. It is designed to hide all details related to the job processing (such as error handling or distribution of tasks across the workers). This allows developers to focus only on enhancing computation issues (in terms of response time, resources utilisation, energy consumption etc.) in their parallel programs rather than parallelism.

2.2.2. Apache Spark

Spark (Zaharia et al., 2010b) is a novel in-memory computing framework written in Scala for Hadoop, proposed in 2010. It was developed to address the problem in the MapReduce model, which accepts only a particular linear data flow format for distributed programs. Spark uses a data structure called Resilient Distributed Dataset (RDD), which is a distributed memory abstraction that allows for in-memory computations on large clusters in a fault-tolerant way (Zaharia et al., 2012). In MapReduce programs, the input data are read from the disk then mapped through a map function, and reduced using a reduce function to get the output data that will be stored on the disk. Whereas in Spark programs, the RDDs serve as a working set for distributed programs, which offer a restricted form of distributed shared memory (Zaharia et al., 2010b). The RDDs support more functions, compared to MapReduce, that can be classified into two main categories; the "transformation" and the "action" functions. The tranformation function can be a map, filter, sample, union, or an intersetcion operation. While an action function can be a reduce, collect, countbykey, take, or takeordered operation. Consequently, the RDDs allow to reduce the latency for both iterative and interactive data analysis applications by several orders of magnitude when compared to Hadoop (Zaharia et al., 2012). Spark is comprised of two main components: a cluster manager and a distributed storage system. Spark supports the native Spark cluster, Hadoop YARN (Liu et al., 2015), or Mesos (Hindman et al., 2011b) as a cluster manager. Also, it supports communication with a multitude of distributed storage systems including HDFS, MapR File System (MapR-FS), and Cassandra (Karpate et al., 2015).

2.2.3. Apache Storm

MapReduce and Hadoop are designed for offline batch processing of static data in cloud environments, which makes them not suitable for processing stream data applications in the cloud (*e.g.*, Twitter) (Peng et al., 2015a). To alleviate this issue, Storm (Peng et al., 2015a) has emerged in 2011 as a promising computation platform for stream data processing. Storm is a distributed computation framework written in Clojure and Java, and designed for performing computations of streams of data in real time. In order to be processed in Storm, an application should be modelled as a directed graph called a topology that includes spouts and bolts, and the data strems of the applications can be routed and grouped through this graph. Particularly, there are different grouping strategies to control the routing of data streams through the directed graph including the field grouping, global grouping, all grouping, and shuffle grouping (Peng et al., 2015a). The spouts are sources of data stream (sequence of tuples), they read data from different sources including database, messaging frameworks, and distributed file systems. The bolts are used to process data messages and to acknowledge the processing of data messages when it is completed. Also, they can be used to generate other data messages for the subsequent bolts to process. Generally, one can utilize the bolts for filtering, managing, aggregating the data messages, or to interact with external systems. Storm can achieve a good reliability by using efficient procedures to control message processing. Also, it has fault-tolerant mechanisms that allow to restart failed workers in case of failures (Xu et al., 2014a).

2.2.4. Task Scheduling

In general, task scheduling is of paramount importance since it aims at allocating a number of dependent and/or independent tasks to the machines having enough resources in the clusters. An effective scheduler can find the optimal task distribution across the machines in a cluster, in accordance with execution time requirements and resources availability. An optimal task distribution minimises the mean execution time of the scheduled tasks and maximises the utilisation of the allocated resources. This is in order to maximise the response time of the received computations (tasks to be processed), and reduce (avoid) resources waste. Each big-data platform in the cloud is equipped with a scheduler that manages the assignment of tasks. Here, we briefly present as examples the well-known schedulers proposed for Hadoop that gained a lot of attention from both industry and academia. In Hadoop, the JobTracker is responsible for scheduling and provisioning the submitted jobs and tasks. It has a scheduling algorithm, which initial implementation was based on the First In First Out (FIFO) principle. The scheduling functions were first regrouped in one daemon. Hadoop developers decided later to subdivide them into one Resource Manager and (per-application) Application Master to ease the addition of new pluggable schedulers. YARN (Yet Another Resources Negotiator) (Liu et al., 2015) is the daemon responsible for managing applications' resources. Facebook and Yahoo! have developed two new schedulers for Hadoop: Fair scheduler (Zaharia et al., 2009) and Capacity scheduler (Raj et al., 2012), respectively.

2.3. Cluster Manager: Apache Mesos

Mesos (Hindman et al., 2011b) is an open-source cluster manager that provides efficient resource usage and sharing across mul-

Table 1
A non-exhaustive list of journals, conferences, and workshops considered in our SLR.

Journals
- Journal of Systems and Software (JSS) - Future Generation Computer Systems (FGCS) - Transactions on Parallel and Distributed Systems (TPDS) - Transactions on Service Computing (TSC)
Conferences - IEEE INFOCOM - IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid) - IEEE International Conference on Distributed Computing Systems (ICDCS)
Workshops - Workshop on Data Engineering - Workshop on Parallel Distributed Processing and Ph.D Forum - Symposium on High-Performance Parallel and Distributed Computing

tiple cluster computing frameworks. It was proposed in 2009 by the University of California, Berkeley. Instead of a centralized approach, Mesos supports a two-level scheduling approach to allocate the resources to the frameworks (in this context, a framework is a software system that executes one or more jobs in a cluster). Hence, Mesos enables efficient resources sharing in a fine-grained way. So, the master node in Mesos decides the amount of resources to be assigned for each framework. Then, each framework accepts the resources it needs and decides which jobs to execute on those resources. This approach can help optimize the allocation of resources as well as provide near-optimal data locality (Hindman et al., 2011b).

3. Methodology of the S.L.R.

The following subsections present our proposed methodology to perform the Systematic Literature Review (SLR), and the outcomes of the SLR:

3.1. Conducting the Study

3.1.1. Data Sources

Following the guidelines given in Kitchenham (2004), we start our SLR using the following relevant search engines: *IEEE Xplore, ACM, Google Scholar, CiteSeer, Engineering Village, Web of Science* and *ScienceDirect*. We perform an electronically-based search and consider the main terms related to this review: *"scheduling", "task scheduling", "scheduler", "MapReduce", "Hadoop", "Spark", "Storm"*, and *"Mesos"*. We use the same search strings for all seven search engines. We look for published scientific literature related to task scheduling in Hadoop, Spark, Storm and Mesos between 2005 and 2016. Then, we restrict our study to a number of journals, conferences, workshops and technical reports having the highest quality and considered as the most important resources in this field. We perform this step by selecting the studies published in journals with high impact factors, conferences/workshops with competitive acceptance rates and technical reports with high number of citations. Also, we check the citation of the studies in order to evaluate their impact in this field. Other studies are rejected for quality reasons (*e.g.*, the study is only a small increment over a previous study, a technical report that is extended into a journal or a conference/workshop paper, etc). Table 1 presents a non-exhaustive list of workshops, conferences and journals considered in our SLR.

3.1.2. Search and Selection Process

The search and selection process for the relevant studies from data sources is organized in three rounds as described in Fig. 3.

- **Round 1:** we perform a mapping study named also a scoping review in order to identify and categorise the primary studies related to the SLR based on their scope. This scoping review helps identify the main issues addressed and studied in the available literature. Next, we select the most relevant studies based on their titles and abstracts. Any irrelevant study is removed. If there is any doubt about any study at this level, the study is kept.
- **Round 2:** it consists of a manual search of the studies obtained in the previous step (*i.e.*, Round 1), which are identified as the main sources for the SLR. It is necessary to check the reliability of the selected studies. To do so, the remaining studies at this step are carefully read. Then, the irrelevant studies are removed based on the selection criteria defined in the work of Dybå and Dingsøyr (2008). More details about the used criteria are given in Section 3.2.
- **Round 3:** we perform a snowball search based on guidelines from Wohlin (2014). We apply a backward snowball search using the reference list of papers obtained in the second round, to identify new papers and studies. We use the same selection criteria (as in *Round 2*) to decide whether to include or exclude a paper. These remaining papers are read carefully.

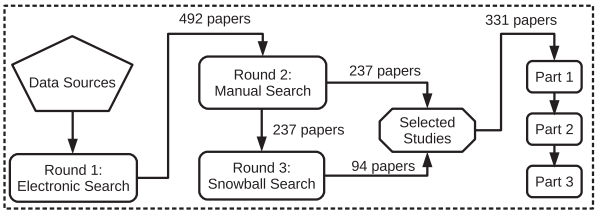

Fig. 3. Overview of SLR Methodology.

3.2. Quality of the Selected Papers

We apply different inclusion and exclusion criteria on the remaining studies in the second and third rounds. These selection criteria can help decide whether to include or not a paper for further search. Only relevant studies that are retained will be used in the SLR analysis to answer our research questions. (1) Only papers describing issues related to Hadoop, Spark, Storm, and Mesos schedulers and proposing models to improve their performance are included. (2) Documents presented in the form of *power point* presentations, abstract, and submitted papers are not included in this review.

3.3. Outcomes of the Study

The different search stages of our SLR identify a total of **586** papers. Specifically, we obtain 492 papers from **Round 1**, from which we extract 237 papers after **Round 2**. Next, we discover 94 new papers during the snowball phase (*i.e.*, **Round 3**). In total, the number of papers analyzed in this SLR is $492 + 94 = 586$ (from Round 1 and Round 3). This is after removing those papers that are not related to task scheduling in Hadoop, Spark, Storm, or Mesos, and duplicates that are found by more than one search engine. When a paper is found by two search engines in Round 1, we keep the one published in the search engine having the highest number of papers. For example, if a paper is found by IEEE and ACM and IEEE has the highest number of obtained study, we keep the one in IEEE and remove the duplicate in ACM. The results obtained on each of the seven search engines, for the three rounds are presented in Table 2.

3.4. SLR Organization

The following paragraphs describe the motivation for each part in the SLR:

Table 2
Results of SLR rounds.

Search Engine	Round 1	Round 2	Round 3
- IEEE	252	143	50
- ACM	95	61	28
- CiteSeer	41	8	2
- Google Scholar	37	6	9
- ScienceDirect	39	12	5
- Web of Science	5	1	0
- Engineering Village	23	6	0
Total	492	237	94

Part 1: *Task Scheduling Issues in Big Data Platforms:*
This part provides a comprehensive overview of task scheduling in Hadoop, Spark, Storm, and Mesos. It aims at identifying the main topics of task scheduling addressed in these frameworks. Hence, it can help determine the challenges and issues that have been studied by both research and commercial communities. The identified challenges and issues will help draw the map of the state of research on task scheduling in these computing frameworks.

Part 2: *Task Scheduling Solutions in Big Data Platforms:*
This part describes the proposed solutions in the existing literature that addressed the scheduling issues identified in ***Part 1***. This exhaustive analysis can help give a comprehensive overview about the characteristics of the proposed solutions, their main objectives and their limitations. In fact, it presents the advantages and limitations of each solution that aimed to improve Hadoop, Spark, Storm, and Mesos schedulers over time. Furthermore, it can help identify some future work that can be addressed by researchers in order to better improve the schedulers of these frameworks.

Part 3: *Research Directions on Task Scheduling in Big Data Platforms:*
This part identifies some of the future work that can be done to cover the drawbacks of the solutions reported in ***Part 2***. Based on the limitations of these proposed solutions, we aim to identify some aspects that can be enhanced to better improve Hadoop, Spark, Storm, and Mesos schedulers. ***Part 3*** draws a roadmap for further studies on task scheduling in these frameworks.

3.5. SLR Analysis Approach

We perform a manual search over the papers found at the different search stages. To do so, we proceed in two steps. First, we skim through the papers, reading the most relevant parts to get an overview of the issues addressed in the existing literature to construct ***Part 1***. Next, we classify the obtained studies in different categories based on their scope to ease the analysis of the selected papers. Also, we compute statistics about the number of published studies and papers *(i)* in each category; and *(ii)* the evolution of this metric over time (from 2005 up to 2016) to get an overview of the most studied scheduling issues in Hadoop, Spark, Storm, and Mesos. Second, we carefully analyse all papers in order to extract the relevant information about the proposed approaches and their limitations to build ***Part 2***. Indeed, we classify the proposed approaches in different categories following the list of issues identified in ***Part 1***. If there is a study that is addressing two issues at the same time, it will be included in both categories. Finally, to develop ***Part 3***, we identify some future work that can be addressed to cover the limitations of the approaches discussed in ***Part 2***. The following sections present and dummyTXdummy- discuss the results of the three parts in our SLR.

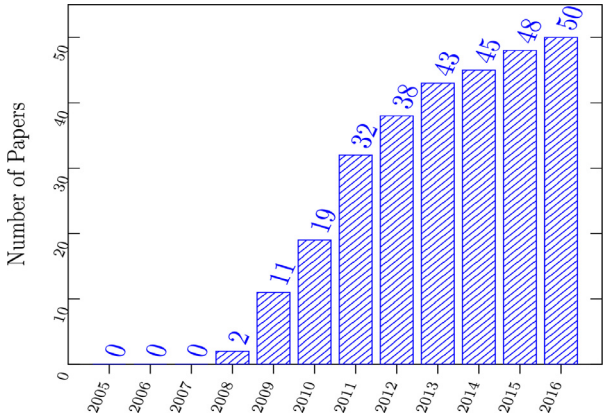

Fig. 4. Distribution of papers over years [2005–2016].

4. Task Scheduling Issues In Big Data Infrastructures

Before addressing the first part of the SLR, we examine the candidate papers based on their publication years to identify the distribution of the related studies over time. Fig. 4 shows the interest of researchers on task scheduling for big data frameworks over time. We notice that during the first three years, after proposing Hadoop in 2005, there was no study that analysed scheduling issues in Hadoop. This is arguably due to the fact that researchers were more interested in the computing functions of Hadoop and were striving to improve them. Next, we observe that in 2008, the topic of scheduling in Hadoop started gaining attraction, with 2 papers published on the topic in 2008. A limited number of studies were performed on the topic between 2009 (11 papers) and 2010 (19 papers). Then, the number of studies significantly increased from 11 papers in 2009 to 50 papers in 2016. This can be explained by the constant increase of the popularity of Hadoop (Hadoop is now widely used by different companies and research labs). Also, the high number of Hadoop users was affecting the overall performance of the scheduler and hence many studies were needed to resolve the issues faced while deploying Hadoop. With the emergence of Mesos (2009), Spark (2010) and Storm (2011), many other works were done to study the performance of these platforms in Cloud environments. We can claim that during three years starting from 2014 until 2016, a minimum of 45 studies were published on Hadoop/Spark/Mesos/Storm scheduling issues each year; highlighting the importance of this research topic. Also, we find that the majority of the studies were addressing scheduling issues in Hadoop and only a few works were analysing the other three platforms. This can be explained by the popularity of Hadoop in both academia and industry and also because Hadoop was proposed before the other platforms (in 2005).

Next, we identify the addressed task scheduling issues in big data platforms using the papers obtained from the three rounds. We find that we can classify these papers into six categories as shown in Fig. 5. The obtained categories can be described as follows:

4.1. Resources Utilisation (65 papers)

In general, the resources allocation process aims to distribute the available resources across the scheduled tasks, in order to ensure a good quality of services for users and to reduce the cost of the services for cloud providers. Particularly, the computing resources (*e.g.*, CPU cores, memory) are distributed using the two basic computing units in a MapReduce job; map and reduce slots. A slot is the basic unit used to abstract the computing resources (e.g., CPU cores, memory) in Hadoop, Spark, and Storm. It is used to

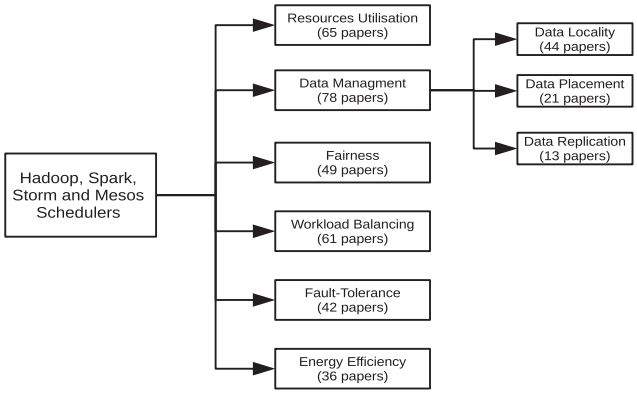

Fig. 5. Scoping review results.

indicate the capacity of a worker (e.g., TaskTracker). There exist two types of slots in a cluster: a slot that can be assigned to a map task, and a slot to be assigned to a reduce task. These computing units are statically configured by the administrator before launching a cluster. The map slots must be only allocated to the map tasks and the reduce slots must be only given to the reduce tasks. The allocation of these slots should be executed under the constraint that the map tasks should be executed and finished before their corresponding reduce task starts. Specifically, the map and reduce phases in a job have a tight dependency on each other. Indeed, the map tasks can be running in parallel since they are small and independent. The reduce tasks will be launched in parallel with the copy and merge phases and will not release their assigned slots until all reduce tasks are completed. Hence, they take much longer time to be finished. So, this fundamental interdependence can lead to a repeatedly observed starvation problem. Furthermore, the map and reduce tasks may have highly varying resources requests over time, which makes it difficult for the scheduler to efficiently utilize the cluster resources. In addition, given the dynamic nature of cloud applications, the resources allocation process can be complex and may fail to allocate the required resources for some jobs (*i.e.*, long running-time jobs) or fail to prevent tasks from the starvation problem. As a result, there can be some straggling map or reduce tasks because of the unexpected contention time for CPU, memory and other resources; resulting in unpredictable execution time for the tasks. Overall, we found 65 studies addressing this issue.

4.2. Data Management (78 papers)

We can claim that the problem of data management in big data platforms can be sub-divided into three main sub-problems as follows:

4.2.1. Data Locality (44 papers)

In big data platforms, the computations (*i.e.*, received workloads) are sent as close as possible to where the input data of the scheduled tasks is located. This is because of the large size of the processed data rather than moving these large data blocks to the computational nodes where the tasks will be running. So, the scheduler decides where to send the computations based on where the data exists. *Data locality* is an important issue addressed by many researchers as shown in Fig. 5. In particular, we find 44 studies that addressed this problem in Hadoop, Spark, Storm and Mesos schedulers. Scheduling tasks based on the locality of their associated data is a crucial problem that can affect the overall performance of a cluster. Indeed, the execution of some jobs or tasks requires the processing of tasks having distributed data across different nodes. Therefore, it is necessary to find a better allocation of these tasks over the available nodes while maximizing the number of tasks executing local data. This is to reduce the total execution time of tasks by reducing the number of non-local-data tasks since these tasks spend more time to read and write data compared to the local-data tasks.

4.2.2. Data Placement (21 papers)

Although, the proposed data locality strategies can help improve the processing of tasks in the nodes having the local input data and enough resources, an extra overhead can be added when processing the non-local data blocks and moving the intermediate data from one node to another to get the final output; which may decrease the overall performance of a cluster. Particularly, the processing of scheduled tasks highly depends on the location of the stored data and their placement strategy. This makes it difficult *(a)* for the platform (*e.g.*, Hadoop, Spark) to distribute the stored data; and *(b)* for the scheduler to assign the scheduled tasks across the available nodes in a cluster. Therefore, there are some studies that are proposed to improve the *data placement* issue within the distributed nodes in these big data platforms in order to improve the strategies responsible for moving the data blocks especially the large data-sets. We find 21 studies that addressed this problem.

4.2.3. Data Replication (13 papers)

Data locality and placement are very important problems as they significantly affect the system performance and the availability of the data. However, some slow or straggling nodes can go down and their data blocks will not be available for sometime, which may affect the overall performance. Therefore, several algorithms are proposed to replicate the data across different processing nodes. For example, if some data cannot be found, the scheduler may find other replicas in other nodes, racks or other clusters and run speculative copies of the straggling tasks. By increasing the number of data replicas, the scheduler may be able to increase the successful processing of tasks that are not able to find their input data. However, the distribution and number of these replica can vary over time due to the dynamic nature of the applications running on a cluster. Therefore, it is necessary to propose better schemes to replicate data over big data platforms nodes. This principle is known as *data replication* and is studied by 13 works as presented in Fig. 5.

4.3. Fairness (49 papers)

Ensuring fairness is particularly important to improve the overall performance of a cluster for big data infrastructures, especially when there are different jobs running concurrently in the cluster. Particularly, 49 papers addressed this issue in the existing literature; which explains the importance of this problem. A job is composed of multiple map and reduce tasks that occupy the available resource slots in the cluster. The slot configuration can differ from one node to another. Also, jobs can have different resource requirements and consumptions over time. For example, some jobs can occupy their assigned resources for long time, more than expected, which may cause the starvation of some other jobs waiting for their turn in the queue. Also, some jobs may take the advantage of occupying the resources to finish their tasks much faster while other jobs can be waiting a long time for their turn to be executed using the same resources. Overall, the scheduler can experience many problems such as jobs starvation and long execution time if it does not define a fair plan to assign the available slots across the jobs and tasks. Consequently, the scheduler should fairly assign the received tasks to the node slots in order to reduce the makespan time between the scheduled tasks. Specifically, it should find the optimal task assignment across the available slots while reducing the overhead generated between the map and reduce tasks to communicate the intermediate results.

4.4. Workload Balancing (61 papers)

The schedulers of cloud platforms, like Hadoop, Spark, Mesos, or Storm, receive multiple jobs and tasks having different characteristics and different resource demands, which may lead to different system workloads. Furthermore, workloads can be imbalanced when jobs are running, under a current scheduler, across the cluster nodes, which may cause important delays and many problems related to the reliability of the system. Therefore, it is very important to balance the workload across the distributed nodes to improve its performance in terms of execution time and resources utilisation; especially for cloud applications that should satisfy some defined response requirements. The workload balancing is very dependent on the fairness allocation of the available slots across the scheduled tasks. In fact, one reason behind unbalanced workload is a bad slot allocation in a worker, which can generate straggling tasks waiting for slots to be released. So, a good slot allocation strategy allows to have a balanced workload. Other factors like resources allocation, type of scheduler, can also affect the workload behavior in each node. Many studies in the available literature (*i.e.*, 61 papers) proposed algorithms to balance the received load in a cluster.

4.5. Fault-Tolerance (42 papers)

Heterogeneity is the norm in cloud environments, where different software configurations can be found. Particularly, Sahoo et al. (2004) claim that the complexity of software systems and applications running on a cluster can cause several software failures (*e.g.*, memory leaks, state corruption), make them prone to bugs and may lead to crashes in the cluster. Physical machines in cloud clusters are subject to failure (*e.g.*, they can be down for some time), which may lead to unexpected delays to process the received jobs. Moreover, big-data platforms' schedulers can experience several task failures because of unpredicted demands of service, hardware outages, loss of input data block, nodes failure, etc. Although, Hadoop, Spark, Mesos and Storm have built fault-tolerance mechanisms to cover the aforementioned limitations, one task failure can cause important delays due to the tight dependency between the map and reduce tasks. Also, it can lead to not-efficient resources utilisation because of an unexpected resources contention. Therefore, many researchers addressed this issue in these platforms. In particular, 42 studies are found in the literature addressing this crucial problem. These studies propose different *Fault-Tolerance* mechanisms to reduce the number of failures in the processing nodes, and hence improve the overall performance of the cluster. In addition, they describe different mechanisms to improve the availability and reliability of platforms components, in order to better improve the availability of the services offered to the users.

4.6. Energy Efficiency (36 papers)

The cost of data-intensive applications running on large clusters represent a critical concern in terms of energy efficiency. For instance, data-intensive applications require more energy when processing the received workload, executing I/O disk operations, and managing and processing the huge amount of data on big data platforms like Hadoop, Spark, or Mesos. Moreover, the design of a scheduler for big data platforms can largely affect the energy consumption of the system on which the applications are executed. For instance, when processing tasks on the nodes where the data exists, the node may receive a large number of tasks and these nodes require more resources to execute them, which can increase the level of energy consumed. Moreover, the nodes in a cloud cluster can experience several failures and can face straggling tasks resulting in more energy being consumed. Therefore, minimizing the energy consumption when processing the received workload is of paramount importance. We find 36 studies that addressed this critical issue in cloud environments. These studies show that there is a trade-off between improving the scheduler performance and the energy consumption on the studied platforms in our SLR.

5. Task Scheduling Solutions In Big Data Infrastructures

After carefully checking the content of the obtained papers, we can group them by scope/objective in order to analyse the solutions proposed to address the issues mentioned in Section 4. The different proposed solutions are described in the following subsections. In the following subsections, we describe each of the proposed approaches in more details. Since the majority of the papers is addressing scheduling issues on Hadoop framework in comparison to the other platforms (including Spark, Storm, and Mesos); this is because of the popularity of Hadoop in both academic and industrial communities. For each category, we first discuss the solutions proposed for Hadoop, and then, if applicable, we discuss solutions proposed for other platforms. Finally, we present a summary to classify and discuss the proposed approaches and the addressed issues.

5.1. Resources Utilisation-aware Scheduling

Although there is a tight dependency between the map and reduce tasks, these two phases are scheduled separately by existing schedulers. Additionally, Zhang et al. (2015) show that the resources consumption varies significantly in these two phases. To mitigate this problem, many studies including (Zhang et al., 2015; Jian et al., 2013c; Liang et al., 2014; PASTORELLI et al., 2015) are proposed to correlate the progress of map and reduce tasks while scheduling them and then assign them slots based on their requirements. Just to name a few, the *Coupling Scheduler* (Jian et al., 2013c), *PRISM* (Zhang et al., 2015), HFSP (PASTORELLI et al., 2015), and *Preedoop* (Liang et al., 2014) are proposed as fine-grained resource-aware MapReduce schedulers for Hadoop. The main goal of these schedulers is to assign available slots according to the variability of the requested resources in each phase of a Hadoop job and the task execution progress. This is in order to reduce the total execution time and to avoid a waste of resources.

Jian et al. (2013c) propose the *Coupling Scheduler*, which is composed of two main parts *(i)* a wait scheduling for the reduce tasks; and *(ii)* a random peeking scheduling for the map tasks. The coupling scheduler can reduce the average job processing time by 21.3% compared to the Fair Scheduler. However, some jobs still face long-waiting times because of other running jobs taking all reduce slots. The idea behind *Preedoop* Liang et al. (2014) is to preempt reduce tasks that are idle and assign their allocated resources to scheduled map tasks, in order to allow for a faster processing at the map phase. This is because reduce tasks waiting for intermediate data, or results from some map tasks, often detain resources that could have been used by some pending map tasks. Liang et al. (2014) report that *Preedoop* can reduce the execution time by up to 66.57%. However, the preemption of the reduce tasks can delay the copy/merge phases of the jobs, which may result in extra delays.

Hadoop Fair Sojourn Protocol (HFSP) (PASTORELLI et al., 2015) is a new scheduling protocol that automatically adapts to resources and workload changes while achieving resources efficiency and short response times for Hadoop. HFSP uses job size and progress information to allocate the available resources to the received tasks. HFSP uses an aging function to make scheduling decisions, such that jobs with higher priority have more chance to get the resources. The priorities of jobs are computed using the aging func-

tion. HFSP can reduce the execution time of the scheduled tasks but, it cannot reduce the number of task failures since it is based on a preemptive technique. Moreover, HFSP delays the scheduling of the map tasks on non-local data, for a certain time and for a fixed number of attempts. This is in order to ensure the processing of the map tasks with local data and improve the performance of Hadoop. Meanwhile, tasks having a lower priority can be processed. However, postponing the processing of the map tasks several times can generate extra delays to the total execution times of the tasks. Hence, some tasks execution times can exceed their specified deadlines; resulting in tasks deadline no-satisfaction.

While processing batch jobs, Hadoop may encounter problems due to inefficient resources utilization, which may generate long and unpredictable delays. Particularly, the static and fixed configuration of slots allocated to the map and reduce tasks in Hadoop can affect their processing time and may lead to a degradation of the cluster resources. To alleviate this issue, some studies, including Wolf et al. (2010); Shanjiang et al. (2013); Liu et al. (2014); Jiayin et al. (2014); Yao et al. (2015); Tang et al. (2016a), introduce the dynamic assignment of the slots to the mappers and reducers depending on their requirements.

For instance, Fair and Efficient slot configuration and Scheduling for Hadoop (*FRESH*) Jiayin et al. (2014) is designed to find the matching between the slot settings and the scheduled tasks to improve the makespan (which is defined as the difference between the longest and the smallest execution time of the running tasks) while guaranteeing a fair distribution of the slots among the map and the reduce tasks. In the same line, Wolf et al. (2010) propose a flexible scheduling allocation scheme called *FLEX* aiming to optimize the response time, deadline satisfaction rates, SLA (Service Level Agreement), and makespan of different type of Hadoop jobs, while allocating the same minimum job slots assigned in the Fair scheduler. Despite the fact that FRESH and FLEX show good performances in terms of total completion time and slot allocation, their proposed scheduling schemes should relax the scheduling decisions in terms of data locality. They only consider how to fairly distribute the slots across the scheduled tasks but, they do not take into account the necessity to schedule them as close as possible to their input data. In addition, they should take into account the remaining execution time of the scheduled jobs to better assign, on the fly, the resources slots.

FiGMR (Mao et al., 2015) is proposed as a fined-grained and dynamic scheduling scheme for Hadoop. FiGMR classifies the nodes in a Hadoop cluster into high or low level performance according to their resources utilisation, and tasks into slow map tasks and slow reduce tasks. It uses historical information from the nodes to dynamically find the tasks that are slowed by a lack of resources. Then, FiGMR launches speculative executions of the slow map and reduce tasks on the high level performance nodes in order to speed up their execution. Overall, FiGMR can reduce the execution time of tasks and improve data locality. But, it requires considerable time to find the slow tasks and to assign them to high level performance nodes, which can result in extra delays to the scheduler.

Because of the large scale of cloud environments, the applications running on top of Hadoop systems are increasingly generating a huge amount of data about the system states (*e.g.*, logfiles, etc). These data can be used to make better scheduling decisions and improve the overall cluster performance. Whereas, the primary Hadoop schedulers rely only on a small amount of information about the Hadoop environment, particularly about the resources allocation/utilisation to make the scheduling decisions. Therefore, many research work (Rasooli and Down, 2012; Zhang et al., 2015; Yao et al., 2014; Rasooli and Down, 2011) have been proposed to build schedulers capable of collecting data about the resources utilisation and adapting their scheduling decisions based on the system states and the events occurring in the cloud computing environment.

For example, *HaSTE* (Yao et al., 2014) is designed as a pluggable scheduler to the existing Hadoop YARN (Liu et al., 2015). HaSTE schedules the received tasks according to many system information like the requested resources and their capacities, and the dependencies between the tasks. It assigns the resources to tasks based on the ratio between the requested resources and the capacity of the available resources. Specifically, HaSTE measures the importance of the received tasks in order to prioritize the most important tasks in a job and to quantify the dependencies between the scheduled tasks. Despite the fact that HaSTE can optimize the resources utilisations, it is limited only to the CPU and memory resources.

Rasooli and Down (2012, 2011) propose to use the collected information about the Hadoop environment to classify the received jobs according to their resources requirements. They implement an algorithm that captures the changes on the system states and adapts the scheduling decisions according to the new system parameters (*e.g.*, queue state, free slots, etc.) in order to reduce the average execution time of the scheduled tasks. But, their proposed approach is associated with an overhead to estimate the execution time of each received job and to make the slot allocation in accordance to the captured system changes.

There are also considerable challenges to scheduling the growing number of tasks with constraints-meeting objectives. Along with the broad deployment of Hadoop schedulers, many studies (Cheng et al., 2015c; Wei et al., 2014; Ullah et al., 2014; Bin et al., 2013; Pletea et al., 2012; Khan et al., 2016) have been proposed to improve the performance of Hadoop in terms of deadline satisfaction. In a nutshell, these schedulers identify the jobs (among the submitted ones) that could be finished within a specific deadline, then, they check the availability of resources to process the jobs. A job will be scheduled if there are enough slots to satisfy its requirements.

Bin et al. (2013) propose a scheduling algorithm that leverages historical information about scheduled and finished tasks and slot performance to make a decision about whether a resources slot (CPU, memory, bandwidth) is good enough for the assigned tasks, the delay threshold, and the tasks' deadlines. Their proposed algorithm also makes a decision about whether a scheduled task should be delayed, since there will be some other available slots better than the selected ones. The proposed scheduler is able to assign the tasks to the suitable slots with acceptable delays. However, delaying small jobs while looking for the most suitable slots can affect their total completion time.

Ullah et al. (2014) consider the remaining execution time of each job when deciding to preempt, in order to maximize the utilization of the slots under the deadline constraints and the execution time requirements. However, they use a static approach to estimate the remaining time, which can affect the average of this value and hence negatively impact the scheduling decisions. Pletea et al. (2012) implement a genetic algorithm to speculate the execution time of the tasks with respect to deadline constraints and the heterogeneity of the distributed resources in a Hadoop cluster. Overall, the genetic-based algorithm must be efficient and fast in terms of execution time while providing the optimal solution to the scheduler. However, the authors do not implement any optimization function to improve the performance of their proposed algorithm; which may negatively impact the performance of the scheduler.

Khan et al. (2016) propose a Hadoop job performance model that can estimate the amount of required resources so that jobs are finished before their deadlines based on the estimation of job completion times. The proposed model uses historical information about job execution records and a Locally Weighted Linear

Regression (LWLR) technique to determine the estimated execution time of Hadoop jobs. It could reduce the total execution time of jobs by up to 95.5% such that jobs are completed before their expected deadlines. However, it only considers the independent Hadoop jobs, which can affect the resource allocation mechanism in Hadoop.

Jiang et al. (2016) claim that the existing scheduler in Spark does not consider any coordination between the utilization of computation and network performance, which may lead to a reduced resource utilisation. Therefore, they design and implement Symbiosis, which is an online scheduler that predicts resources imbalance in Spark cluster. Jiang et al. (2016) propose to schedule computation-intensive tasks (with data locality) and network-intensive tasks (without data locality) on the same CPU core in Symbiosis. When several tasks are scheduled and competing on the same CPU core, they integrate a priority-based strategy to select which task to process first. Symbiosis is able to reduce the total completion times of Spark jobs by 11.9% when compared to the current scheduler of Spark framework. However, the authors do not consider the resource and network utilisation for the intermediate steps that involve especially network transfers hence, it can add extra delays to the processing of the jobs.

Apache Storm is the most popular stream processing system used in industry. It uses the default round-robin scheduling strategy, which does not consider the resources availability and demand. To alleviate this issue, R-Storm Peng et al. (2015b) is proposed to satisfy soft and hard resources constraints, minimize the network latency, and increase the overall throughput. Peng et al. implement a scheduling algorithm using the Quadratic Knapsack Problem (QKP) and find that R-Storm outperforms Storm in terms of throughput (30%-47%) and CPU utilisation (69%-350%).

Mesos (Hindman et al., 2011a) possesses a fined-grained resource sharing scheduler that controls the sharing of resources across the applications running on the platform. In other words, Mesos decides the amount of resources that can be assigned to an application, and this application decides the tasks to run on them. This approach allows the application to communicate with the available resources to build a scalable and efficient platform. Consequently, Mesos can achieve better resource utilisation and near-optimal data locality. But, it does not take into consideration the requirements of applications running on Mesos while assigning them the resources, which may result in a waste of resources.

5.2. Data Management-aware Scheduling

Data management is a hot issue that caught the attention of many researchers. This is because of its direct impact on the performance of big data platforms including those of the task scheduling techniques. For instance, the performance of big data platforms' schedulers is highly dependent on the procedures dedicated to managing the data to be processed in the computing nodes. This issue is extensively studied by researchers who aim to efficiently distribute data schemes across the nodes. In the following paragraphs, we describe different approaches proposed by researchers to improve the ***data locality, data placement*** and ***data replication*** schemes.

5.2.1. Data Locality-aware Scheduling

MapReduce is widely used in many systems where there are dynamic changes over time. However, it lacks the flexibility to support small and incremental data changes. To cover this limitation, the *IncMR* framework (Cairong et al., 2012) is proposed to improve the data locality incrementally. IncMR fetches the prior state of runs in the system and combines it with the newly added data. This can help find better scheduling decisions according to the new changes in the systems. So, the state of system runs periodically get updated for future incremental data changes. The conducted experiments in Cairong et al. (2012) show that their approach has a good impact on non-iterative applications running in MapReduce. Indeed, the running time is faster than the one obtained when processing the entire input data. But, IncMR is subject to high network utilisation and the large size of files storing the system states consumes resources. Therefore, it is very important to optimize the storage of the states of the system (in terms of size and location) in order to get efficient processing times and optimize the network bandwidth.

Tseng-Yi et al. (2013) propose the Locality-Aware Scheduling Algorithm (*LASA*) in order to achieve better resource assignments and data locality in Hadoop schedulers. They present a mathematical model to calculate the weight of data interference that will be given to LASA. The data interference is derived using the number of data in each node having free slots. Next, LASA selects the node having the smallest weight and data locality to process the received tasks. But, LASA does not guarantee a fair distribution of the workload across the Hadoop nodes.

Kao and Chen (2016) present a real-time scheduling framework for Hadoop that can guarantee data locality for interactive applications. In this work, the authors present both a scheduler and a dispatcher for Hadoop. The scheduler is responsible for assigning tasks when the required resources are available, and the dispatcher considers the data locality of these tasks. The performance of the proposed framework is evaluated using synthesized workload and it shows good results in terms of execution time and energy consumption optimization. Whereas it does not consider the priority of the tasks while assigning the tasks to the nodes having their local block data; which can affect the performance of the applications running on Hadoop.

Zaharia et al. (2008, 2010a) present the Longest Approximate Time End (*LATE*) algorithm, which collects data about the running tasks and assigns weights to tasks based on their progress. Using historical information about the weights assigned to tasks in the past, LATE prioritizes the new tasks waiting to be executed. LATE predicts the finish times of each task and speculates on the ones that can meet most the response time in the future. The proposed algorithm can improve the response time of the schedulers by a factor of 2.

Later, Liying et al. (2011) extended LATE by introducing a delay on the processing of tasks. Each task being delayed for a maximum of K times. They propose that a task should wait for T/S seconds before checking the availability of slots in the nodes having local data. In this equation, T is the average task execution time and S is the number of slots in the cluster. Since a task could be delayed up to K times, it is possible to have some tasks waiting for up to K * T/S seconds before being processed. Although, their proposed algorithm can reduce the overall response time of tasks and improve the system throughput, it has to sort twice the whole system to find the tasks having local input data and the task that will be launched speculatively; which may add extra delays to the response time. Moreover, the value of *K* should be suitable for the system status, to avoid the task starvation problem and system performance degradation. Also, LATE faces some issues (*i.e.*, inaccurate estimation of the remaining time of tasks) in calculating the task progress and identifying the straggling ones due to its static approach. In addition, it does not distinguish between the map and reduce tasks while calculating their progress, which may affect its performance.

Processing data within a requesting node for a data-intensive application represents a key factor to improve the scheduling performance in Hadoop. Many researchers (*e.g.*, Zaharia et al. (2008); Hui et al. (2012b); Guo et al. (2012); Xue et al. (2015)) have been working extensively to solve this problem by evaluating the impact of many factors on the data locality. This can help identify the

correlation between data locality and those identified factors and hence schedule tasks on the processing nodes as close as possible to their input data. As illustration for this solution, the research work in Hui et al. (2012b); Guo et al. (2012) describe mathematical models and scheduling algorithms to evaluate the impact of many configuration factors on data locality. Examples of the configuration parameters can be the input data size and type, the dependency between the data input of tasks, the number of nodes, the network traffic, etc. They propose to perform a job grouping step before scheduling the tasks; the jobs belonging to the same group should be ordered based on their priority and the locality of their input data. Also, they propose to schedule multiple tasks simultaneously instead of one by one to consider the impact of other tasks that may not guarantee better scheduling's performance. These proposed algorithms can increase the number of tasks processed using local data-blocks, which can reduce their execution time. However, these solutions do not show a good improvement when job sizes are large. This is because large jobs have more distributed input data across different nodes and hence the proposed algorithms cannot guarantee to have a maximum number of local-data tasks for these jobs. The proposed approaches work well only when the job sizes are small.

Although, Hadoop and Spark are characterized by a good performance when allocating the resources and processing the received workload, they show a poor performance in handling skewed data. For instance, scheduled tasks can experience several types of failure, because of straggling tasks and skewed data. To solve this problematic issue, many studies are proposed to avoid data skewness and to find the optimal distribution of data (*e.g.*, Liu et al. (2014); Liroz-Gistau et al. (2016); Coppa and Finocchi (2015); Zheng et al. (2014)). For example, FP-Hadoop Liroz-Gistau et al. (2016) is a framework that tackles the problem of data skewness for the reduce tasks, by integrating a new phase in Hadoop job processing. The intermediate phase is called intermediate reduce (IR). The IR can process intermediate values between the map and reduce tasks in parallel with other tasks. This approach can help speedup the processing of the intermediate values even when all of them are associated with the same key. The experimental results show that FP-Hadoop has a better performance compared to Hadoop and can help reduce the execution times of reduce tasks by a factor of 10 and the total execution time of jobs by a factor of 5. But, Hadoop jobs can experience extra delays when there are no skewed data, because the IR workers add more time to the total execution time of a job.

Although, the data locality issue is tackled as one problem in the studies presented above, other research works address it separately for map and reduce tasks, as described in the sequel.

Data Locality of Map Tasks

The pre-fetching techniques of the input data are very important to improve the data locality factor for the map tasks and avoid the data skewness problem. Particularly, there are several research work that address this issue including Sangwon et al. (2009); Chunguang et al. (2013); Wang et al. (2013). They propose prefetching and scheduling techniques to address the data locality of map tasks. These two techniques look for the suitable candidate input data for the map tasks. Also, they select which reducer is better in order to minimize the network traffic required to shuffle the key-value pairs. Although these techniques can improve the data locality of the map tasks, they cannot balance the load across the processing nodes. This is because the proposed techniques can only improve the number of local map tasks and does not take into account the resources utilisation and load balancing.

Asahara et al. (2012) propose *LoadAtomizer* to improve the data locality of map tasks and minimize the completion time of multiple jobs. The LoadAtomizer strategy consists in assigning tasks on lightly loaded storage with consideration to data locality, which can balance the load between the storage nodes. LoadAtomizer can avoid I/O congestion and reduce the CPU I/O waiting time ratio of the map tasks. It could reduce the total execution time of jobs by up to 18.6%. However, it cannot reduce the data skewness for the map and reduce tasks since it aims at balancing the I/O load and increase the data locality of the scheduled tasks.

In Xiaohong et al. (2011b); Polo et al. (2013), the authors present scheduling techniques to improve the data locality of map tasks by dynamically collecting information about the received workload. Also, they propose to dynamically control and adjust the process responsible for allocating the slots across the received jobs to meet their specified deadlines. The obtained results show that the proposed algorithm gives a better performance in terms of the amount of transmitted data across the network and the execution time. To better improve these proposed scheduling techniques, the authors may consider the different types of received tasks (short, long, continuous, etc.) and calculate the remaining execution time of the scheduled tasks.

Data Locality of Reduce Tasks

There are a few other research work Hammoud and Sakr (2011); Jian et al. (2013b) that are proposed to improve the data-locality for the reduce tasks. Hammoud and Sakr (2011) proposed the Locality Aware Reduce Task Scheduling (*LARTS*) algorithm to maximize data locality for the reduce tasks, *i.e.*, the intermediate results generated by the mappers. LARTS uses an early shuffling technique to minimise the overall execution time by activating the reduce task after a defined percentage of mappers commit (*e.g.*, a default value of 5%). Therefore, it can help avoid data skewness and reduce the scheduling delay between the mappers and reducers. LARTS is based on locating the sweet spots of the reducers. These sweet spots can be defined as the time during which a reducer can recognize all its partitions. These spots are located by LARTS statically. Therefore, dynamic identifications of these spots can improve the performance of LARTS. Jian et al. (2013b) propose a stochastic optimization framework to improve the data locality of reducers and minimize the cost associated with fetching the intermediate data. However, this approach works under a fixed number of map and reduce slots in Hadoop; which may lead to an under or over utilization of the available resources.

Motivated by the challenges associated with the default scheduler in Storm, Xu et al. (2014b) proposed a new stream-processing framework T-Storm based on the Storm framework. In fact, Storm uses a default scheduler that assigns the received workload based on the round-robin algorithm without considering the data locality factor. Also, Storm assigns the workload to the nodes regardless of their requirements or the availability of the resources Xu et al. (2014b). Hence, T-Storm is proposed to use run time states to dynamically assign tasks to the nodes where the data are located so that none of the workers is overloaded or underloaded, which could accelerate the task processing and minimize the online traffic in Storm. Moreover, it can achieve better performance with a smaller number of nodes since it allows fine-grained control over the nodes consolidation. The experimental analysis shows that T-Storm can achieve a better performance (up to 84% speedup), and a better data locality for the stream processing applications. Although T-STorm can achieve a good performance with 30% less worker nodes, T-Storm still lacks a fault-tolerance mechanism to handle failures in these nodes which have to process more workload than others.

5.2.2. Data Placement-aware Scheduling

Many studies are proposed to improve data placement strategies within Hadoop and provide optimized data placement schemes, *e.g.*, Jiong et al. (2010); Xiaohong et al. (2011a);

Sharma et al. (2013). These optimized schemes can help improve the data locality for the scheduled tasks.

For instance, Jiong et al. (2010) proposed to adapt the data placement schemes in accordance to the workload distribution in Hadoop clusters. They introduce an algorithm to initially distribute input data across the nodes in accordance to the node's data processing speed. Second, they describe a data redistribution algorithm to dynamically solve the data skew issue, by reorganizing file fragments through the cluster nodes based on their computing ratios. Although these proposed algorithms can help improve the placement and the locality of data in Hadoop clusters, they do not include a mechanism to handle redundant file fragments, neither do they provide a mechanism to redistribute dynamically the data for data-intensive applications working together.

A Hierarchical MapReduce scheduler called *HybridMR* is presented in Sharma et al. (2013) to classify the received MapReduce jobs based on their expected overhead to guide the placement between the physical and virtual machines in Hadoop clusters. In addition, *HybridMR* can dynamically organize the resources orchestration between the different map and reduce tasks and hence decrease their processing time by 40% over a virtual cluster and save around 43% of energy. Despite the fact that *HybridMR* shows a good performance, it cannot handle different types of workload in heterogeneous Hadoop environments and ensure a balanced workload between the nodes.

MRA++ (Anjos et al., 2015) is a new Mapreduce framework for Hadoop proposed to handle large heterogeneous clusters. It allows Hadoop to efficiently process data-intensive applications. This is by training tasks to collect information about data distribution in order to dynamically update the data placement schemes within the framework. MRA++ is mainly composed of a data division module responsible for dividing the data for the tasks, a task scheduling module that controls the task assignment to the available slots, a clustering control module, that controls task execution, and a measuring task module that controls and distributes the data. MRA++ can improve performance of Hadoop by 66.73%. It can also reduce the network traffic by more than 70% in 10 Mbps networks. But, it adds extra delays to the tasks' processing times since they are collecting more information and have to wait for the measuring task module to assign them to the appropriate nodes.

5.2.3. Data Replication-aware Scheduling

Several studies address the problem of data replication in Hadoop to improve storage space utilization, *e.g.*, Hui et al. (2012a); Ananthanarayanan et al. (2011); Abad et al. (2011). For instance, Hui et al. (2012a) propose the Availability-Aware MapReduce Data Placement (*ADAPT*) algorithm to optimally dispatch data across the nodes according to their availability, to reduce network traffic without increasing the number of data replica. Their strategy can improve network traffic, however, it may lead to more disk utilization.

Ananthanarayanan et al. (2011) propose *Scarlett*, which uses a proactive replication scheme that periodically replicates files based on the predicted popularity of data. In other words, Scarlett calculates a replication factor for the data based on their observed usage probability in the past history in order to avoid the problem of data skewness. Scarlett is an off-line system that improves data replicas using a proactive approach but, many changes can occur in a Hadoop storage system including recurrent as well as nonrecurrent changes.

While Scarlett uses a proactive approach, Abad et al. (2011) present the Distributed Adaptive Data REplication (*DARE*) algorithm, which is a reactive approach to adapt the data popularity changes at smaller time scales. The DARE algorithm aims at determining how many replicas to allocate; and at controlling where to place them using a probabilistic sampling and competitive ageing algorithm. As a result, the data locality factor in DARE is improved by 7 times when compared to the FIFO scheduler and by 85% in comparison to the Fair scheduler. However, both Scarlett and DARE do not take into account data with low replica factors.

5.3. Fairness-aware Scheduling

In big data platforms' clusters, data locality and fairness represent two conflicting challenges. Indeed, to achieve a good data locality, a maximum number of tasks should be submitted close to their computation data. However, to achieve fairness, resources should be allocated to the tasks after being requested in order to reduce tasks delays (Zaharia et al., 2010a). Many research work including Jiayin et al. (2014); Isard et al. (2009); Yin et al. (2013); Hui et al. (2012c); Phuong et al. (2012); Cho et al. (2013) are proposed in the available literature to solve the above issues.

Jiayin et al. (2014) present FaiR and Efficient slot configuration and Scheduling algorithm for Hadoop (*FRESH*), to find the matching between the submitted tasks and the available slots. FRESH can help not only minimize the makespan but, also fairly assign available resources across the scheduled tasks. In Hadoop, each node has a specific number of slots. However, the Hadoop scheduler continuously receives concurrent jobs that require different slots configurations. Therefore, Jiayin et al. (2014) extend FRESH by adding a new management plan to dynamically find the best slot setting. In other words, FRESH allows to dynamically change the assignment of slots between the map and reduce tasks according to the availability of slots and the requirement of the tasks. After a slot finishes its assigned task, FRESH can assign it to another task. While FRESH can improve the assignment of slots and the fairness of the distribution of resources among the scheduled tasks, it does not ensure a better memory usage.

Isard et al. (2009) propose Quincy, which is a flexible and efficient scheduling algorithm to compute the scheduling distribution among the different nodes with a min-cost flow while improving data locality, fairness and starvation freedom factors. However, Quincy is only formulated based on the number of computers in a cluster and there is no effort to dynamically reduce its cost in terms of data transfer.

Yin et al. (2013) show that processor-based schedulers like the Fair scheduler can lead to a degradation of performance in terms of execution time, in a multi-user environment. Therefore, they propose the Hybrid Parallel pessimistic Fair Schedule Protocol (*H-PFSP*), which is able to finish jobs later than the Fair Scheduler and improve the mean flow time of jobs while improving the fairness between the tasks and jobs. The *H-PFSP* use information about the finished tasks over time to estimate the remaining execution time of the scheduled jobs at predefined intervals and make incremental estimations updates. The *H-PFSP* can reduce the total execution time but, it cannot guarantee an efficient resources utilisation in the cluster.

Hui et al. (2012c) describe a scheduling algorithm based on a multi-queue task planning to adjust the maximum number of tasks assigned to each node by defining the value of fairness threshold "*K%*". The K%-Fairness scheduling algorithm can be suitable for different types of workloads in MapReduce to achieve maximum of data locality under this constraint. However, this approach cannot support much continuous/dependent jobs in the queue since it cannot decide how to fairly distribute them (due to tight dependencies between them) and reduce the associated overhead while processing them.

Phuong et al. (2012) propose a HyBrid-Scheduling (*HyBS*) algorithm for Hadoop. It is dedicated for processing data-intensive workloads based on the dynamic priority and data locality of the scheduled tasks. In other words, it uses dynamic priorities information, estimated map running times, and service level values

defined by the user to minimize the delays for concurrent running tasks which may have different lengths. *HyBS* can guarantee a fair distribution between the map and reduce tasks. Also, it decreases the waiting time between the map and reduce tasks by resolving data dependencies for data intensive MapReduce workloads. This is by assigning a dynamic priority, obtained from historical Hadoop log files, to the different tasks received in order to reduce the latency for different length (in terms of execution time) concurrent jobs. *HyBS* is using a greedy fractional Knapsack algorithm (Phuong et al., 2012) to assign jobs to the appropriate processing nodes.

The authors of Cho et al. (2013) propose *Natjam* to evaluate the smart eviction policies for jobs and tasks, the priorities for real time job scheduling and the resources availability and usage. Natjam is based on two main priorities policies. These policies are based on the remaining time of each task: Shortest Remaining Time (SRT) in which tasks characterized by the shortest remaining time are the candidate to be suspended; and Longest Remaining Time (LRT) in which tasks characterized by the longest remaining time will be suspended. The two proposed policies that are based on priorities aim to reduce the execution time of each task. Next, they propose *Natjam-R*, a generalization of *Natjam*, which specifies hard and fix deadlines for jobs and tasks Cho et al. (2013). So, the deadline of Hadoop jobs can automatically define the priority of the jobs and their composing tasks for accessing the resources slots. This approach was found to have a negative impact (*i.e.*, delay) on short running tasks that have low priorities, since they can get evicted several times.

Guo et al. (2016b) present FlexSlot a task slot management scheme for Hadoop schedulers that can identify the straggling map tasks and adjust their assigned slots accordingly. This approach can accelerate the execution of these straggling tasks and avoid extra delays. FlexSlot changes the number of slots on each node in Hadoop according to the collected information about resource utilisation and the straggling map tasks. Hence, the available resources in Hadoop cluster are efficiently utilised and the problem of data skew can be mitigated with an adaptive speculative execution strategy. The obtained results show that FlexSlot could reduce the total job completion times by up to 47.2% compared to the Hadoop scheduler. However, FlexSlot generates a delay that can impact the processing of the Hadoop job since it is using the task-killing-based approach in the slot memory resizing. In addition, FlexSlot allows to kill tasks multiple times, which may generate not only extra delays but also may cause the failure of the whole job.

5.4. Workload Balancing-aware Scheduling

Distributing the received loads across computing nodes represents a crucial problem in big data platforms' systems. An efficient distribution can help improve the resources utilisation and guarantee a fair distribution of tasks to be processed, resulting in a better performance for their schedulers.

For instance, Chao et al. (2009) report that the First Come First Served (FCFS) strategy works well only for jobs belonging to the same class (*e.g.*, having the same size, the same resources requirements). Thus, they propose a *Triple-Queue Scheduler*, which dynamically classifies the received Hadoop jobs into three different categories based on their expected CPU and I/O utilisation. Also, they integrate a workload prediction mechanism called *MR-Predict*, which determines the type of the workloads on the fly and distributes them fairly (based on their type) across the different queues. MR-Predict can increase the map tasks throughput by up to 30% and reduce the total makespan by 20% over the Triple-Queue scheduler. However, it still faces other issues to efficiently manage the resources utilisation; to reduce the resources waste and to improve the data locality.

Mao et al. (2011) propose a load-driven Dynamic Slot Controller (*DSC*) algorithm that can adjust the slots of map and reduce tasks according to the workload of the slave nodes. Hence, DSC can improve the CPU utilisation by 34% and the response time by 17% when processing 10 GB of data. But, the DSC algorithm does not take into account the issue of data locality while balancing the load between the nodes.

Fei et al. (2013) propose Shortest Period Scheduler (*SPS*) to ensure that most of the jobs are finished before their specified deadlines. SPS supports preemption and can make dynamic decisions when new workflow plans are received periodically in the scheduler. SPS is limited to scheduling the independent tasks within the received workflow. However, it should cover dependent tasks and analyse the impact of the communication between the scheduled tasks on their expected deadline and the resources utilisation.

Peng et al. (2012) propose a Workload Characteristic Oriented (*WCO*) scheduler to consider the characteristics of the running workloads and make smart decisions that can improve the resource utilisation. The WCO scheduler is able to dynamically detect the difference between the received and the running workloads. Consequently, it can help balance the CPU and I/O usage among Hadoop nodes which could improve the system throughput by 17%. WCO can be improved by enhancing its static analysis method used for the workload characteristics.

Cheng et al. (2014) proposed to use the configuration of large MapReduce workloads to design a self-adaptive task scheduling approach. Their proposed solution consists of an Ant-based algorithm that allows for an efficient workload distribution across the available resources, based on the tasks characteristics. As a result, their approach can improve the average completion time of the scheduled tasks by 11%. Also, they find that their proposed Ant-based algorithm is more suitable for large jobs that have multiple rounds of map task execution. However, this proposed algorithm cannot cover multi-tenant scenarios in MapReduce. In addition, Cheng et al. Cheng et al. (2014) do not provide details about the optimization of the Ant-algorithm to reduce its execution overhead.

Tang et al. (2016b) propose a scheduling algorithm (to optimize the workflow scheduling) in which jobs are represented as Directed Acyclic Graph (DAG) and classified into I/O intensive or computations intensive jobs. Then, the scheduler can assign priorities to the jobs based on their types and assign the available slots with respect to data locality and load balancing. But, this proposed approach was found to work well only for large jobs. It can negatively impact the performance of small jobs.

Li et al. (2014) propose WOrkflow over HAdoop (*WOHA*) to improve workflow deadline satisfaction rates in Hadoop clusters. WOHA relies on the job ordering and progress requirements to select the worklflow that falls furthest from its progress based on the Longest Path First and Highest Level First algorithms. As a result, WOHA can improve workflow deadline satisfaction rates in Hadoop clusters by 10% compared to the existing scheduling solutions (FIFO, Fair and Capacity schedulers). WOHA uses the workloads received over time to estimate the deadline of each task that are not known by the scheduler ahead of time. In addition, the dynamic nature of Hadoop workloads may affect the performance of the scheduler. But, developers of WOHA do not include these two criteria while implementing it.

Rasooli and Down (2012) propose a hybrid solution to select the appropriate scheduling approach to use based on the number of the incoming jobs and the available resources. This proposed solution is a combination of three different schedulers: FIFO, Fair sharing and Classification, and Optimization based Scheduler for Heterogeneous Hadoop (*COSHH*). The COSHH scheduler uses Linear

Programming (LP) to classify the incoming workloads and find an efficient resources allocation using job classes requirements. The aim of this hybrid scheduler is to improve the average completion time, fairness, locality and scheduling times in order to improve Hadoop's scheduling performance. The FIFO algorithm is used for under-loaded systems, the Fair Sharing algorithm is used when the system is balanced and the COSHH is used when the system is overloaded (*i.e.*, peak hours). Rasooli et al. define three different usage scenarios and specify when to use each of them, however, they do not provide thresholds that can be used to decide about which scheduler to follow.

Sidhanta et al. (2016) propose OptEx, which is a closed-form model that analytically analyses and estimates the job completion time on Spark. OptEx model uses the size of input dataset, the number of nodes within the cluster and the number of operations in the job to be scheduled, to estimate the total completion time of the given job. The results show that it can estimate the job completion time in Spark with a mean relative error of 6% when integrated with the scheduler of Spark. Furthermore, OptEx can estimate the optimal cost for running a Spark job under a specific deadline in the Service Level Objective (SLO) with an accuracy of 98%. Although OptEx is the first model in the open literature to analytically estimate the job completion time on Spark, it only considers the job profiles of PageRank and WordCount as parameters along with the size of the cluster and the dataset. This model cannot be representative for real cluster where different workload having different profiles are running.

Sparrow (Ousterhout et al., 2013) is a distributed scheduler that allows the machines to operate autonomously and support more requests from different applications running Hadoop or Spark. In other words, Sparrow is a decentralized scheduler across a set of machines that operate together to accommodate additional workload from users. When a scheduler in a machine fails, other machines may accept its received requests and process it according to their availability. The challenge in Sparrow consists in balancing the load between the machines' schedulers and providing shorter response times, especially when the distributed schedulers make conflicting scheduling decisions. Sparrow uses three main techniques to improve the performance of its schedulers: *Batch Sampling, Late Binding*, and *Policies and Constraints. Batch Sampling* schedules *m* tasks in a job on the lightly loaded machines, rather than scheduling them one by one. *Late Binding* places the *m* tasks on the machine queue only when it is ready to accept new tasks to reduce the waiting time in the queue that is based on FCFS. The *Policies and Constraints* are to control the scheduling decisions and avoid the conflicts on the scheduling decisions. Sparrow allows to distribute the received workload and balance it across the available workers in a shorter time. While Sparrow can reduce the execution time of the jobs by up to 40%, it lacks mechanisms to take into account the requirements of the received workload while distributing them across the nodes. Also, it does not consider the resources availability on each node, the schedulers accept the new requests if the queue is not yet empty, which can overload the machines. Moreover in case of a scheduler failure, the meta-data scheduling of the tasks running on that machine will not be shared with the other machines.

5.5. Fault-Tolerance-aware Scheduling

Although, Hadoop, Spark, Storm, and Mesos are equipped with some built-in fault-tolerance mechanisms, they still experience several tasks failures due to unforeseen events in the Cloud. For example, the HDFS in Hadoop keeps multiple replicas of data blocks on different machines to ensure an effective data restoration in case of a node failure. The failed map and reduce tasks will be rescheduled on other nodes and re-executed from scratch. This fault-tolerant solution is associated with a high cost because of the task re-execution events, which can significantly affect the performance of the Hadoop scheduler. To address the aforementioned limitations, researchers have proposed new mechanisms to improve the fault-tolerance of Hadoop (*e.g.*, Quiane-Ruiz et al. (2011); Yuan and Wang (2013)).

Quiane-Ruiz et al. (2011) proposed Recovery Algorithm for Fast-Tracking (*RAFT*) for Hadoop to dynamically save the states of tasks at regular intervals and at different stages. This approach allows the JobTracker to restart the tasks from the last checkpoint in the event of a failure. Indeed, RAFT enables the Hadoop scheduler to not re-execute the finished tasks of the failed jobs since their intermediate data are saved. So, the scheduler will only re-execute the failed tasks. As a result of this strategy, RAFT can reduce the total execution time of tasks by 23% under different failure scenarios.

Yuan and Wang (2013) propose an approach that dynamically detects the failures of scheduled tasks and makes backups of the tasks. In case of a failure, the scheduler would launch the failed tasks on other nodes without losing their intermediate data. Although the two works presented in Quiane-Ruiz et al. (2011); Yuan and Wang (2013) can improve the fault-tolerance of the system, they do not provide a mechanism to improve the availability of the checkpoints and the used backups.

Xu et al. (2012) claim that the long delays of jobs are due to the straggling tasks and that the LATE scheduler Zaharia et al. (2008) can make inaccurate estimations of the remaining time of tasks, which may lead to resource waste. Thus, they propose a dynamic tuning algorithm that uses historical information about tasks progresses to tune the weights of each map and reduce tasks. In addition, they design an evaluation approach that decides whether to launch a straggling task on another node when there are free slots in order to reduce the execution time and resources waste. However, they do not propose a mechanism to distinguish between different types of straggling tasks, *i.e.*, whether it is a map or a reduce task. This is particularly important since it can affect the speculative executions.

Dinu and Ng (2012) analyse the behavior of the Hadoop framework under different types of failures and report that the recovery time of the failed components in Hadoop can be long and can cause important delays, which may affect the overall performance of a cluster. They claim that sharing information about straggling and failed tasks between JobTrackers and TaskTrackers, can significantly improve the success rate of task executions.

To quickly detect Hadoop nodes failures, Hao and Haopeng (2011) develop an adaptive heartbeat interval module for the JobTracker. Using this module, the JobTracker can dynamically estimate its *expiry interval* for various job sizes. They show that when the expiry interval decreases (which means that the average number of heartbeats sent to the JobTracker increases), the total execution time of small jobs decreases. In addition, they propose a reputation-based detector to evaluate the reputation of the workers. A worker will be marked as failed when its reputation is lower than a threshold. They claim that if equipped with their proposed tools, Hadoop can detect node failures in shorter times and balance the load received by the JobTracker to reduce job execution times. However, they only consider the job size when deciding to adjust the heartbeat interval and they do not include other parameters related to the nodes environment (*e.g.*, running load, availability of resources, failure occurrence).

In addition to the above work, *Astro* (Gupta et al., 2014) is designed to predict anomalies in Hadoop clusters and identify the most important metrics contributing towards the failure of the scheduled tasks using different machine learning algorithms. The predictive model in Astro can detect anomalies in systems early and send a feedback to the scheduler. These early notifications

can improve resources usage by 64.23% compared to existing implementations of Hadoop schedulers. Astro can be improved by adding mechanisms that enable a better distribution of workloads between the nodes of the cluster. This would reduce the execution time of the scheduled tasks by 26.68% during the time of an anomaly.

The execution of MapReduce jobs in Hadoop clusters can undergo many failures or other issues, which may affect the response time and delay submitted jobs. Preemtion is proposed as an effective solution to identify straggling jobs and tasks in advance and make quick scheduling decisions to prevent a waste of resources. Different approaches based on speculative executions have been proposed to address this issue in distributed Hadoop clusters:

Qi et al. (2014), develop an algorithm called Maximum Cost Performance (*MCP*), to improve existing speculative execution strategies. However, MCP was found to negatively impact the scheduling time of some jobs (batch jobs in particular) Shanjiang et al. (2014).

The Combination Re-Execution Scheduling Technology (*CREST*) Lei et al. (2011) algorithm is proposed to improve MCP, by considering data locality during the speculative scheduling of slow running tasks. The authors propose to optimally re-execute map tasks having local data instead of launching speculative tasks without considering data locality. However, there is a cost associated with the replication of executed map tasks.

Self-Adaptive MapReduce scheduling (*SAMR*) uses hardware system information over time to estimate the progress of the tasks and adjust the weights of the map and reduce tasks, in order to minimize the total completion time (Quan et al., 2010). However, SAMR does not consider jobs characteristics in terms of size, execution time, weights, etc.

Enhanced Self-Adaptive MapReduce scheduling (*ESAMR*) is designed to overcome the drawbacks of SAMR and consider system information about straggling tasks, jobs length, etc. ESAMR uses the K-means clustering algorithm to estimate tasks execution times and identify slow tasks. It is more accurate than SAMR and LATE (Zaharia et al., 2008). Although ESAMR can identify straggling map and reduce tasks and improve the execution time of jobs, it does not provide rescheduling mechanisms for these straggling tasks and does not improve the number of the finished tasks.

AdapTive faiLure-Aware Scheduler (*ATLAS*) Soualhia et al. (2015) is proposed as a new scheduler for Hadoop that adapts its scheduling decisions to events occurring in the cloud environment. ATLAS can identify task failures in advance and adjust its scheduling decisions on the fly based on statistical models. It can reduce task failure rates, resources utilisation and total execution time. However, it requires training its predictive model at fixed time intervals, which may negatively impact the scheduling time. Also, it may face problems to find the appropriate scheduling rule or it can give wrong predictions that can cause the failure of tasks.

Yildiz et al. (2015, 2017) propose *Chronos*, a failure-aware scheduling strategy that enables early actions to recover the failed tasks in Hadoop. Chronos is characterized by a pre-emption technique to carefully allocate resources to the recovered tasks. It can reduce the job completion times by up to 55%. However, it is still relying on wait and kill pre-emptive strategies, which can lead to resource waste and degrade the performance of Hadoop clusters.

5.6. Energy Efficiency-aware Scheduling

The total energy consumption of the applications running on big data platforms depends on many factors including the number of the low-load nodes and the processed load on each node. Several studies addressed the issue of finding good task assignments while saving the energy, *e.g.*, Mashayekhy et al. (2015); Wen (2016); Paraskevopoulos and Gounaris (2011); Chen et al. (2012).

Mashayekhy et al. (2015) propose to model the problem of saving energy on MapReduce jobs as an integer programming problem and design two heuristics Energy-MapReduce Scheduling Algorithm I and II (EMRSA-I and EMRSA-II). The proposed model considers the dependencies between the reduce tasks and the map tasks such that all tasks are finished before their expected deadlines, while the main goal of the proposed approach is to minimize the amount of energy consumed by these map and reduce tasks. EMRSA-I and EMRSA-II are evaluated using TeraSort, PageRank, and K-means clustering applications and are able to reduce the energy consumption by up to 40% on average, when compared to the default scheduler of Hadoop. In addition, they can reduce the makespan between the processed MapReduce jobs. However, in the proposed model, the authors assume that map tasks belonging to the same job should all receive resources slots (same assumption for the reduce tasks), before the execution of the job. However, this is not generally the case in Hadoop and such restriction can delay the execution of a MapReduce job if even a single map or reduce task fail to obtain a slot.

Wen (2016) propose a dynamic task assignment approach to reduce the overall system energy consumption for dynamic Cloud Hosts (CHs). The idea of the approach is to have a set of power-on/suspending thresholds to satisfy the constant and variable traffic loads, migration overhead, and the processing power between the CHs. Based on the proposed thresholds, the Hadoop scheduler can dynamically assign tasks to satisfy those constraints and achieve better energy efficiency. The evaluation of these schemes shows that setting the thresholds between the CHs can help obtain the lowest energy consumption and acceptable execution times for Hadoop jobs. However, there is an overhead that comes when suspending or powering on the CHs, which can affect the network traffic in Hadoop, especially when the frequency of these two operations is high.

Paraskevopoulos and Gounaris (2011) propose a strategy to schedule the tasks for Hadoop, while balancing between energy consumption and response time. Their proposed strategy can help identify the nodes in a cluster that can satisfy the constraint of less energy in a reasonable response time. Next, it can determine which nodes should be turned on–or–off, and when that should be done, based on the derived nodes and the received workload in the cluster. The experimental results show a significant improvement on energy consumption without sacrificing the scheduler performance. In this work, the authors only consider the response times of the jobs when deciding about the best task scheduling policies that can minimize energy consumption. They do not consider balancing the workload between the nodes that have the highest impact on the energy consumption of a Hadoop cluster.

Chen et al. (2012) introduce a scheduling approach for Hadoop to minimize the energy consumption of MapReduce jobs. The proposed approach consists in dividing the jobs into time-sensitive jobs and less time-sensitive ones. The former group of jobs are run on dedicated nodes where there are enough resources, while the later ones run on the remaining nodes in the cluster. Chen et al. (2012) introduce a management framework for Hadoop named Berkeley Energy Efficient MapReduce (BEEMR). BEEMR is able to reduce the energy consumption in Hadoop clusters by 40–50% under tight design constraints. However, although BEEMR can achieve energy savings for workloads with significant interactive analysis, BEEMR cannot reduce the processing times of long running jobs that have inherently low levels of parallelism, even when all resources in the cluster were available Chen et al. (2012).

5.7. Discussion

In the previous subsections, we describe the different types of scheduling approaches available in the open literature to solve the

issues presented in Section 4. Table 3 presents a classification of these approaches and the addressed issues. The main addressed issues in the studied papers are: improve the resources utilisation, reduce tasks delays, tasks dependency consideration, reduce the execution times of tasks and jobs, improve the deadline satisfaction of tasks, reduce network traffic, improve the data locality/placement/replication strategies, reduce the data skew, balance the workload, reduce failures rates (tasks, workers, etc), and reduce the amount of energy consumed in big data platforms. The proposed approaches can be classified into three main categories: *dynamic, constrained*, and *adaptive*. We observe that most of the existing solutions propose to collect and use data about the environment where the computations are processed (*e.g.*, clusters, machines, workers). This can be explained by the dynamic behavior and structure of the cloud computing environment where Hadoop, Spark, Storm, and Mesos platforms are deployed. This requires to adapt the scheduling decisions of these platforms according to the continuous changes across the clusters.

In general, we observe the lack of formal description of the addressed issues and the proposed solutions in the papers analysed in this SLR. Indeed, we notice that most papers conduct empirical studies (*e.g.*, Wolf et al. (2010); Shanjiang et al. (2013); Liu et al. (2014)) and very few work propose analytical models (*e.g.*, Convolbo et al. (2016); Guo et al. (2014); Cheng et al. (2015a)) to solve the scheduling issues. So, an interesting direction could be to improve the empirical studies by developing formal models in order to improve the performance of the Hadoop, Spark, Storm, and Mesos' schedulers. Another concern is the benchmarks (*e.g.*, WordCount, TeraSort) used to implement and build the proposed solutions. The use of these benchmarks is highly dependent on the objective of the study (*e.g.*, resources optimization, failures recovery). The absence of dataset to configure and define the parameters of these benchmarks may lead to biased results. Furthermore, we find that several studies conducted by the academics do not become commercialized and part of Apache projects (Hadoop, Spark, and Storm) or Mesos. Finally, we can conclude that applying and adapting the proposed solutions for Hadoop to Spark, Storm, and Mesos could be an interesting direction since we notice that only a few work are done to improve the performance of Spark, Storm, and Mesos compared to Hadoop.

6. Research Directions on Task Scheduling in Big Data Infrastructures

In this section, we present some suggestions for potential research directions using the results from the above paragraphs. These suggestions can help build a roadmap for future work related to task scheduling in the studied platforms, *i.e.*, Hadoop, Spark, Storm, and Mesos.

6.1. Resources Utilisation-aware Scheduling

During our study, we observe that existing work in the literature propose different approaches to assign the map and reduce slots and evaluate the performance of the scheduler. Moreover, most of the studies randomly select the map tasks that satisfy the slots requirements. But, it is very important to include the data locality issue while scheduling the map tasks in order to improve their execution and hence, avoid the data skewness problem and reduce the execution times of the reduce tasks (as mentioned in Section 5). Besides, we notice that the task preemption while occupying or waiting for a slot can cause unpredictable delays. So, an efficient approach is required to manage task preemption in a way that do not generate an overhead and avoid task starvation. Furthermore, analysing several factors (*e.g.*, queue state, available slots, received workload, number of nodes) on the resources utilisation can be helpful to guide the scheduler to change its scheduling decisions based on the events occurring in its environment. The scheduler may consider different constraints-objectives along with a fair distribution of load constraint while scheduling the tasks. Examples of these constraints can be the Service Level Agreement (SLA), the number of tasks having local data, the transferred data in the network, etc. Finally, more studies need to be done in order to improve the performance of Spark, Mesos, and Storm schedulers in terms of resources utilisation since we find very few work done in this aspect. For instance, considering the characteristics of workload running on Spark, Mesos, and Storm can guide the scheduler to make better scheduling decisions while assigning the available resources to the tasks. Also, we believe that some of the existing solutions proposed to improve the performance of Hadoop-Mapreduce can be reused and adapted for the other platforms. To do that, one should consider the data structure/format and the way these data are processed within these platforms. For example, new approaches should consider the characteristics of the in-memory operations performed in Spark using the Resilient Distributed Dataset (RDD). The allocation of resources to tasks in Spark should be done, taking into account the amount of memory required to store the intermediate data between the tasks, since a Spark job does the whole computation and then stores the final output to the disk. In the case of Storm, it is important to consider the structure of the applications processing the spouts and the bolts, and the dependency between these tasks, especially the bolts that process the data read from input streams or the generated output of other bolt tasks. This is because there might be some bolt tasks running in sequential or/and parallel. Therefore, one should consider the parallel and the sequential aspects while scheduling the tasks in Storm. For Mesos, the type of frameworks should be considered (*e.g.* CPU-intensive, memory-intensive or network-intensive frameworks), when adapting existing solutions to improve the resources utilisation. This is because the type of the framework can affect the performance of the assigned resources by Mesos. So, Mesos should consider not only the amount of the assigned resources to each framework but also the type of resources to offer them. In addition, the two levels of the scheduling process of Mesos require synchronization between them.

6.2. Data Management-aware Scheduling

While reviewing the data-management aware scheduling solutions in the literature, we notice that the proposed schemes that place or duplicate the data across the cluster nodes are not made based on a workload analysis. Therefore, analysing the impact of different workloads, data placements and replication mechanisms are needed to improve the data locality of the scheduled tasks. Moreover, having a large number of local tasks may cause an unbalanced workload across the processing nodes. Hierarchical scheduling can be a solution for this issue; this approach consists in having one scheduling layer to consider the data locality and another scheduling layer to handle workload balancing. These two layers should communicate their global and local information to cooperate together. In addition, we observe that most of the solutions that try to achieve data locality are characterized by an overhead due to the cost of finding the optimal distribution of file fragments. This would significantly affect the performance of the scheduler. So, better solutions need to be developed to reduce this overhead. On the other hand, existing scheduling solutions cannot guarantee high data locality for large jobs since there is a lot of data to be transferred. Therefore, efficient approaches should be developed in this direction in order to handle different job scales. Moreover, we notice that the majority of the studies we find are related to Hadoop schedulers. So, more work should be done to

Table 3
An overview of task scheduling issues and solutions in big data platforms.

Issue/ Solution	Map/Reduce Dependency	Slot Assignment Assignment	Data Collecting	Load profiling	Prefetchning & Shuffling Techniques	Recovery Techniques	Failure Prediction
Resources Variability	Zhang et al. (2015) Jian et al. (2013c) Liang et al. (2014) PASTORELLI et al. (2015)	Wolf et al. (2010) Shanjiang et al. (2013) Liu et al. (2014) Jiayin et al. (2014) Yao et al. (2015) Tang et al. (2016a) Isard et al. (2009)	Jiang et al. (2016) Hindman et al. (2011a)				
Task Delays		Wolf et al. (2010) Shanjiang et al. (2013) Liu et al. (2014) Jiayin et al. (2014) Yao et al. (2015) Tang et al. (2016a) Isard et al. (2009)	Isard et al. (2009) Hui et al. (2012c) Phuong et al. (2012) Cho et al. (2013)				
Dependency Tasks	Yao et al. (2014) Peng et al. (2015b)		Yao et al. (2014) Peng et al. (2015b)				
Execution Time		Sidhanta et al. (2016)	Rasooli and Down (2012) Zhang et al. (2015) Yao et al. (2014) Rasooli and Down (2011)				
Deadline Satisfaction			Cheng et al. (2015c) Wei et al. (2014) Ullah et al. (2014) Bin et al. (2013) Pletea et al. (2012) Khan et al. (2016) Peng et al. (2015b)				
Data Locality/ Network Traffic	Zaharia et al. (2008) Zaharia et al. (2010a) Liying et al. (2011)		Peng et al. (2015b) Hindman et al. (2011a) Cairong et al. (2012) Tseng-Yi et al. (2013) Kao and Chen (2016) Zaharia et al. (2008) Zaharia et al. (2010a) Liying et al. (2011) Xu et al. (2014b)	Zaharia et al. (2008) Hui et al. (2012b) Guo et al. (2012) Xue et al. (2015)			
Data Skew	Liu et al. (2014) Liroz-Gistau et al. (2016) Coppa and Finocchi (2015) Zheng et al. (2014)	Guo et al. (2016b)	Xiaohong et al. (2011b) Polo et al. (2013)	Asahara et al. (2012)	Sangwon et al. (2009) Chunguang et al. (2013) Wang et al. (2013) Hammoud and Sakr (2011) Jian et al. (2013b)		
Data Placement			Jiong et al. (2010) Xiaohong et al. (2011a) Sharma et al. (2013)				
Data Replication			Hui et al. (2012a) Ananthanarayanan et al. (2011) Abad et al. (2011)				
Unbalanced Workload			Fei et al. (2013) Cheng et al. (2014) Li et al. (2014) Ousterhout et al. (2013)	Chao et al. (2009) Mao et al. (2011) Peng et al. (2012) Tang et al. (2016b) Rasooli and Down (2012)			
Failures Rates			Quiane-Ruiz et al. (2011) Yuan and Wang (2013) Lei et al. (2011) Xu et al. (2012) Dinu and Ng (2012) Hao and Haopeng (2011) Qi et al. (2014) Quan et al. (2010)			Quiane-Ruiz et al. (2011) Yuan and Wang (2013) Lei et al. (2011) Xu et al. (2012) Dinu and Ng (2012) Hao and Haopeng (2011) Qi et al. (2014) Quan et al. (2010)	Gupta et al. (2014) Soualhia et al. (2015) Yildiz et al. (2015) Yildiz et al. (2017)
Energy Consumption	Mashayekhy et al. (2015) Wen (2016)	Mashayekhy et al. (2015) Wen (2016)	Paraskevopoulos and Gounaris (2011)	Chen et al. (2012)			

analyse the performance of Spark, Mesos and Storm in terms of data locality, replication and duplication.

6.3. Fairness-aware Scheduling

Distributing the available resources slots among the scheduled tasks is important in order to avoid the starvation problem. However, to the best of our knowledge, the research studies that address this issue do not consider the difference between the map and reduce tasks during slots assignments. Indeed, while estimating the remaining execution time, the proposed solutions do not distinguish between them; which may affect the slots assignment since the map and reduce tasks have different slot requirements. In addition, developing a constraint-solver that combines several objectives including data locality, resources utilisation and fairness can significantly improve the performance of schedulers on these platforms. Implementing heuristics can solve this constrained-problem, however, it may cause an overhead. Moreover, we notice that most of the studies in this research direction do not handle fairness for continuous jobs that can occupy the available slots for longer time than the small ones. Therefore, proposing an efficient approach that can estimate the amount of slots required to guarantee successful processing for both continuous and non-continuous jobs while having a fair distribution for the available slots would be useful. Also, it is very important to reduce the amount of data communication for dependent jobs, in order to mitigate the overhead due to the transfer of intermediate data.

6.4. Workload Balancing-aware Scheduling

To improve the performance of schedulers of the studied platforms, one can develop efficient solutions that estimate the remaining execution time of the scheduled tasks based on their progress rate in order to redistribute the loads across the nodes. Indeed, existing schedulers use a simple approach to estimate the remaining execution time and hence the resulting average execution time cannot be used in heterogeneous clusters and may lead to unbalanced workloads. Besides, the performance of the prediction models used to estimate the type of received workloads can significantly affect the load distribution. Hence, it is required to build robust models with high accuracy, to predict the characteristics of the upcoming loads. Based on these predictions, the scheduler can make better decisions when assigning the slots and guarantee data locality for the scheduled tasks. Moreover, it is very important to propose a model to adjust the scheduling decisions considering dependencies between the tasks within the workload. This can help reduce the overhead to communicate the intermediate results and allow for faster processing. Also, it can reduce the amount of data transferred in the network. The analysis of the impact of virtual machines placements on the processing of the load could enable a better workload distribution.

6.5. Fault-Tolerance-aware Scheduling

Reducing the occurrence of failures in big data platforms is very important in order to improve the resources utilisation and the performance of the scheduled tasks. However, existing schedulers only make use of a limited amount of information when re-executing failed tasks. This is due to the lack of information sharing between the different components of these frameworks. Adaptive solutions that collect data about the events occurring in the cloud environment and adjust the decisions of the scheduler accordingly could help avoid decisions leading to task failures. Moreover, the speculative execution still experiences many failures and waste of resources due to inaccurate estimations of the scheduled tasks progresses or the availability of resources. This can affect their starting time and the number of speculatively executed tasks. Therefore, it is very important to analyse the impact of different factors on the start time and the number of speculative executions required for the straggling tasks. Finally, it is very important to distinguish between the failure of a map and a reduce task since they have different impacts on the processing of tasks.

6.6. Energy-Efficiency-aware Scheduling

Determining the configuration for big data platforms like Hadoop or Spark can be very helpful to achieve energy savings and make efficient scheduling decisions. Also, analysing the correlation between the number of nodes in a cluster and the amount of energy consumed can be relevant in order to specify the nodes to turn on–or–off, so that the number of active nodes satisfy the requirements of the received tasks. Moreover, analysing the level of parallelism in a cluster when the number of active nodes increases can be an important direction to guide the scheduler to scale up–or–down the level of parallelism for the scheduled tasks especially for the long jobs. In addition, most of the studies proposed to improve the performance of Hadoop schedulers does not consider the impact of delaying the execution of tasks on the overall performance of the scheduler in terms of users' requirements. Another aspect that can be interesting for future studies is to analyse the impact of frequencies at which machines are turned on or off in a cluster, especially large ones, on the amount of energy consumed. Although turning off some nodes in a cluster can help reduce energy consumption, this can generate more traffic on the network since the scheduled tasks may not find their data on the nodes where they will be executed; which could increase the number of data transfer in the cluster.

7. Conclusion

In recent years, task scheduling has evolved to become a critical factor that can significantly affect the performance of cloud frameworks such as Hadoop, Spark, Storm and Mesos. This crucial issue is addressed by many researchers. However, to the best of our knowledge, there is no extensive study on the literature of task scheduling for these frameworks that classifies and discusses the proposed approaches. Hence, we perform a SLR to review existing literature related to this topic. In this work, we review 586 papers and identify the most important factors affecting the performance of the proposed schedulers. We discuss these factors in general with their associated challenges and issues namely, resources utilisation, total execution time, energy efficiency etc. Moreover, we categorize the existing scheduling approaches from the literature (*e.g.*, adaptive, constrained, dynamic, multi-objective) and summarise their benefits and limitations. Our mapping study allows us to classify the scheduling issues in different categories including resources management, data management (data locality, data placement and data replication), fairness, workload balancing, fault-tolerance, and energy efficiency. We describe and discuss the approaches proposed to address these issues, classifying them into four main groups; dynamic scheduling approaches, constrained scheduling approaches, and adaptive scheduling approaches. Finally, we outline some directions for future research that can be included in a roadmap for research on task and jobs scheduling in Hadoop, Spark, Storm and Mesos frameworks.

References

Abad, C., Yi, L., Campbell, R., 2011. DARE: adaptive data replication for efficient cluster scheduling. In: Proceedings of International Conference on Cluster Computing, pp. 159–168.

Ananthanarayanan, G., Agarwal, S., Kandula, S., Greenberg, A., Stoica, I., Harlan, D., Harris, E., 2011. Scarlett: coping with skewed content popularity in MapReduce clusters. In: Proceedings of Conference on Computer Systems, pp. 287–300.
Anjos, J.C., Carrera, I., Kolberg, W., Tibola, A.L., Arantes, L.B., Geyer, C.R., 2015. Mra++: scheduling and data placement on mapreduce for heterogeneous environments. Future Gener. Comput. Syst. 42, 22–35.
Asahara, M., Nakadai, S., Araki, T., 2012. LoadAtomizer: a locality and I/O load aware task scheduler for MapReduce. In: Proceedings of IEEE International Conference onCloud Computing Technology and Science, pp. 317–324.
Bin, Y., Xiaoshe, D., Pengfei, Z., Zhengdong, Z., Qiang, L., Zhe, W., 2013. A delay scheduling algorithm based on history time in heterogeneous environments. In: Proceedings of ChinaGrid Annual Conference, pp. 86–91.
Cairong, Y., Xin, Y., Ze, Y., Min, L., Xiaolin, L., 2012. IncMR: incremental data processing based on MapReduce. In: Proceedings of International Conference on Cloud Computing, pp. 534–541.
Chao, T., Haojie, Z., Yongqiang, H., Li, Z., 2009. A dynamic MapReduce scheduler for heterogeneous workloads. In: Proceedings of International Conference on Grid and Cooperative Computing, pp. 218–224.
Chen, Y., Alspaugh, S., Borthakur, D., Katz, R., 2012. Energy efficiency for large-scale MapReduce workloads with significant interactive analysis. In: Proceedings of the ACM European Conference on Computer Systems, pp. 43–56.
Cheng, D., Lama, P., Jiang, C., Zhou, X., 2015. Towards energy efficiency in heterogeneous Hadoop clusters by adaptive task assignment. In: IEEE International Conference on Distributed Computing Systems, pp. 359–368.
Cheng, D., Rao, J., Guo, Y., Zhou, X., 2014. Improving MapReduce performance in heterogeneous environments with adaptive task tuning. In: Proceedings of International Middleware Conference, pp. 97–108.
Cheng, D., Rao, J., Jiang, C., Zhou, X., 2015. Resource and deadline-aware Job scheduling in dynamic Hadoop clusters. In: IEEE International Parallel and Distributed Processing Symposium, pp. 956–965.
Cho, B., Rahman, M., Chajed, T., Gupta, I., Abad, C., Roberts, N., Lin, P., 2013. Natjam: design and evaluation of eviction policies for supporting priorities and deadlines in Mapreduce clusters. In: Proceedings of Annual Symposium on Cloud Computing, pp. 6:1–6:17.
Chunguang, W., Qingbo, W., Yusong, T., Wenzhu, W., Quanyuan, W., 2013. Locality based data partitioning in MapReduce. In: Proceedings of International Conference on Computational Science and Engineering, pp. 1310–1317.
Convolbo, M.W., Chou, J., Lu, S., Chung, Y.C., 2016. DRASH: a data replication-aware scheduler in geo-distributed data centers. In: IEEE International Conference on Cloud Computing Technology and Science, pp. 302–309.
Coppa, E., Finocchi, I., 2015. On data skewness, stragglers, and MapReduce progress indicators. In: Proceedings of ACM Symposium on Cloud Computing, pp. 139–152.
Dean, J., Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters. ACM Commun. 51 (1), 107–113.
Dinu, F., Ng, T., 2012. Understanding the effects and implications of compute node related failures in Hadoop. In: Proceedings of International Symposium on High--Performance Parallel and Distributed Computing, pp. 187–198.
Dybå, T., Dingsøyr, T., 2008. Empirical studies of agile software development: a systematic review. Inf. Softw. Technol. 50 (9–10), 833–859.
Fei, T., Hao, Y., Tianrui, L., Yan, Y., Zhao, L., 2013. Scheduling real-time workflow on MapReduce-based cloud. In: Proceedings of International Conference on Innovative Computing Technology, pp. 117–122.
Guo, Y., Rao, J., Jiang, C., Zhou, X., 2014. FlexSlot: moving Hadoop into the cloud with flexible slot management. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, pp. 959–969.
Guo, Y., Rao, J., Jiang, C., Zhou, X., 2016. Moving MapReduce into the cloud with flexible slot management and speculative execution. IEEE Trans. Parallel Distrib. Syst. PP (99), 1–14.
Guo, Z., Fox, G., Zhou, M., 2012. Investigation of data locality in MapReduce. In: Proceedings of International Symposium on Cluster, Cloud and Grid Computing, pp. 419–426.
Gupta, C., Bansal, M., Chuang, T.-C., Sinha, R., Ben-romdhane, S., 2014. Astro: a predictive model for anomaly detection and feedback-based scheduling on Hadoop. In: Proceedings of International Conference on Big Data, pp. 854–862.
Apache Hadoop Project. 2017. http://hadoop.apache.org/.
Hammoud, M., Sakr, M., 2011. Locality-aware reduce task scheduling for MapReduce. In: Proceedings of International Conference on Cloud Computing Technology and Science, pp. 570–576.
Hao, Z., Haopeng, C., 2011. Adaptive failure detection via heartbeat under Hadoop. In: Proceedings of IEEE Asia-Pacific Services Computing Conference, pp. 231–238.
Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A., Katz, R., Shenker, S., Stoica, I., 2011. Mesos: a platform for fine-grained resource sharing in the data center. In: Proceedings of Conference on Networked Systems Design and Implementation, pp. 295–308.
Hindman, B., Konwinski, A., Zaharia, M., Ghodsi, A., Joseph, A.D., Katz, R., Shenker, S., Stoica, I., 2011. Mesos: a platform for fine-grained resource sharing in the data center. In: Proceedings of USENIX Conference on Networked Systems Design and Implementation, pp. 295–308.
Hui, J., Xi, Y., Xian-He, S., Raicu, I., 2012. ADAPT: availability-aware MapReduce data placement for non-dedicated distributed computing. In: Proceedings of International Conference on Distributed Computing Systems, pp. 516–525.
Hui, Z., Shuqiang, Y., Zhikun, C., Hong, Y., Songchang, J., 2012. An locality-aware scheduling based on a novel scheduling model to improve system throughput of MapReduce Cluster. In: Proceedings of International Conference on Computer Science and Network Technology, pp. 111–115.
Hui, Z., Shuqiang, Y., Zhikun, C., Hua, F., Jinghu, X., 2012. K%-Fair scheduling: a flexible task scheduling strategy for balancing fairness and efficiency in MapReduce systems. In: Proceedings of International Conference on Computer Science and Network Technology, pp. 629–633.
Isard, M., Prabhakaran, V., Currey, J., Wieder, U., Talwar, K., Goldberg, A., 2009. Quincy: fair scheduling for distributed computing clusters. In: Proceedings of ACM SIGOPS Symposium on Operating Systems Principles, pp. 261–276.
Jian, T., Shicong, M., Xiaoqiao, M., Li, Z., 2013. Improving reduce task data locality for sequential MapReduce jobs. In: Proceedings of IEEE INFOCOM, pp. 1627–1635.
Jian, T., Shicong, M., Xiaoqiao, M., Li, Z., 2013. Improving reduce task data locality for sequential MapReduce jobs. In: Proceding of IEEE INFOCOM, pp. 1627–1635.
Jian, T., Xiaoqiao, M., Li, Z., 2013. Coupling task progress for MapReduce resource-aware scheduling. In: Proceedings of IEEE INFOCOM, pp. 1618–1626.
Jiang, J., Ma, S., Li, B., Li, B., 2016. Symbiosis: network-aware task scheduling in data-parallel frameworks. In: IEEE International Conference on Computer Communications, pp. 1–9.
Jiayin, W., Yi, Y., Ying, M., Bo, S., Ningfang, M., 2014. FRESH: fair and efficient slot configuration and scheduling for Hadoop clusters. In: International Conference on Cloud Computing, pp. 761–768.
Jiong, X., Shu, Y., Xiaojun, R., Zhiyang, D., Yun, T., Majors, J., Manzanares, A., Xiao, Q., 2010. Improving MapReduce performance through data placement in heterogeneous Hadoop clusters. In: Proceedings of International Symposium on Parallel Distributed Processing, pp. 1–9.
Kao, Y.-C., Chen, Y.-S., 2016. Data-locality-aware mapreduce real-time scheduling framework. J. Syst. Softw. 112, 65–77.
Karpate, S., Joshi, A., Dosani, J., Abraham, J., 2015. Cascket: a binary protocol based C client-driver for apache cassandra. In: Proceedings of International Conference on Advances in Computing, Communications and Informatics, pp. 387–393.
Khan, M., Jin, Y., Li, M., Xiang, Y., Jiang, C., 2016. Hadoop performance modeling for job estimation and resource provisioning. IEEE Trans. Parallel Distrib. Syst. 27 (2), 441–454.
Kitchenham, B., 2004. Procedure for performing systemic reviews. Technical Report. Keele University and NICTA, Australia.
Kurazumi, S., Tsumura, T., Saito, S., Matsuo, H., 2012. Dynamic processing slots scheduling for I/O intensive jobs of Hadoop MapReduce. In: Proceedings of International Conference on Networking and Computing, pp. 288–292.
Lee, K., Lee, Y., Choi, H., Chung, Y., Moon, B., 2012. Parallel data processing with MapReduce: a survey. SIGMOD Rec. 40 (4), 11–20.
Lei, L., Tianyu, W., Chunming, H., 2011. CREST: towards fast speculation of straggler tasks in MapReduce. In: Proceedings of International Conference on e-Business Engineering, pp. 311–316.
Li, S., Hu, S., Wang, S., Su, L., Abdelzaher, T., Gupta, I., Pace, R., 2014. WOHA: deadline-aware Map-Reduce workflow scheduling framework over Hadoop clusters. In: Proceedings of International Conference on Distributed Computing Systems, pp. 93–103.
Liang, Y., Wang, Y., Fan, M., Zhang, C., Zhu, Y., 2014. Predoop: preempting reduce task for job execution accelerations. In: Big Data Benchmarks, Performance Optimization, and Emerging Hardware. Springer, pp. 167–180.
Liroz-Gistau, M., Akbarinia, R., Agrawal, D., Valduriez, P., 2016. Fp-hadoop: efficient processing of skewed MapReduce jobs. Inf. Syst. 60, 69–84.
Liu, N., Yang, X., Sun, X.H., Jenkins, J., Ross, R., 2015. YARNsim: simulating Hadoop YARN. In: Proceedings of IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 637–646.
Liu, Z., Zhang, Q., Ahmed, R., Boutaba, R., Liu, Y., Gong, Z., 2014. Dynamic resource allocation for MapReduce with partitioning skew. IEEE Trans. Comput. 13 (9), 1–14.
Liying, L., Zhuo, T., Renfa, L., Liu, Y., 2011. New improvement of the Hadoop relevant data locality scheduling algorithm based on LATE. In: Proceedings of International Conference on Mechatronic Science, Electric Engineering and Computer, pp. 1419–1422.
Mao, H., Hu, S., Zhang, Z., Xiao, L., Ruan, L., 2011. A load-driven task scheduler with adaptive DSC for MapReduce. In: Proceedings of International Conference on Green Computing and Communications, pp. 28–33.
Mao, Y., Zhong, H., Wang, L., 2015. A fine-grained and dynamic MapReduce task scheduling scheme for the heterogeneous cloud environment. In: Proceedings of International Symposium on Distributed Computing and Applications for Business Engineering and Science, pp. 155–158.
Mashayekhy, L., Nejad, M.M., Grosu, D., Zhang, Q., Shi, W., 2015. Energy-aware scheduling of MapReduce jobs for big data applications. IEEE Trans. Parallel Distrib. Syst. 26 (10), 2720–2733.
Ousterhout, K., Wendell, P., Zaharia, M., Stoica, I., 2013. Sparrow: distributed, low latency scheduling. In: Proceedings of ACM Symposium on Operating Systems Principles, pp. 69–84.
Paraskevopoulos, P., Gounaris, A., 2011. Optimal tradeoff between energy consumption and response time in large-scale MapReduce clusters. In: Proceedings of Panhellenic Conference on Informatics, pp. 144–148.
PASTORELLI, M., Carra, D., Dell'Amico, M., Michiardi, P., 2015. HFSP: bringing size-based scheduling to Hadoop. IEEE Trans. Cloud Comput. PP (99), 1–14.

Patil, S., Deshmukh, S., 2012. Survey on task assignment techniques in Hadoop. Int. J. Comput. Appl. 59 (14), 15–18.
Peng, B., Hosseini, M., Hong, Z., Farivar, R., Campbell, R., 2015. R-Storm: resource-aware scheduling in storm. In: Proceedings of Annual Middleware Conference, pp. 149–161.
Peng, B., Hosseini, M., Hong, Z., Farivar, R., Campbell, R., 2015. R-Storm: resource-aware scheduling in storm. In: Proceedings of Annual Middleware Conference, pp. 149–161.
Peng, L., Young, C.L., Chen, W., Bing, B., Junliang, C., Zomaya, A., 2012. Workload characteristic oriented scheduler for MapReduce. In: Proceedings of International Conference on Parallel and Distributed Systems, pp. 156–163.
Phuong, N., Simon, T., Halem, M., Chapman, D., Le, Q., 2012. A hybrid scheduling algorithm for data intensive workloads in a MapReduce environment. In: Proceedings of International Conference on Utility and Cloud Computing, pp. 161–167.
Pletea, D., Pop, F., Cristea, V., 2012. Speculative genetic scheduling method for Hadoop environments. In: Proceedings of International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, pp. 281–286.
Polo, J., Becerra, Y., Carrera, D., Steinder, M., Whalley, I., Torres, J., Ayguade, E., 2013. Deadline-based MapReduce workload management. IEEE Trans. Netw. Serv. Manage. 10 (2), 231–244.
Qi, C., Cheng, L., Zhen, X., 2014. Improving MapReduce performance using smart speculative execution strategy. IEEE Trans. Comput. 63 (4), 954–967.
Quan, C., Daqiang, Z., Minyi, G., Qianni, D., Song, G., 2010. SAMR: a self-adaptive MapReduce scheduling algorithm in heterogeneous environment. In: Proceedings of International Conference on Computer and Information Technology, pp. 2736–2743.
Quiane-Ruiz, J.-A., Pinkel, C., Schad, J., Dittrich, J., 2011. RAFTing MapReduce: fast recovery on the RAFT. In: Proceedings of International Conference on Data Engineering, pp. 589–600.
Raj, A., Kaur, K., Dutta, U., Sandeep, V., Rao, S., 2012. Enhancement of Hadoop clusters with virtualization using the capacity scheduler. In: Proceedings of International Conference on Services in Emerging Markets, pp. 50–57.
Rao, B., Reddy, L., 2012. Survey on improved scheduling in Hadoop MapReduce in cloud environments. CoRR abs/1207.0780.
Rasooli, A., Down, D., 2011. An adaptive scheduling algorithm for dynamic heterogeneous Hadoop systems. In: Conference of the Center for Advanced Studies on Collaborative Research, pp. 30–44.
Rasooli, A., Down, D., 2012. A hybrid scheduling approach for scalable heterogeneous Hadoop systems. In: Proceedings of IEEE Conference on High Performance Computing, Networking Storage and Analysis, pp. 1284–1291.
Sahoo, R.K., Squillante, M.S., Sivasubramaniam, A., Zhang, Y., 2004. Failure data analysis of a large-scale heterogeneous server environment. In: International Conference on Dependable Systems and Networks, pp. 772–781.
Sangwon, S., Ingook, J., Kyungchang, W., Inkyo, K., Jin-Soo, K., Seungryoul, M., 2009. HPMR: prefetching and pre-shuffling in shared MapReduce computation environment. In: Proceedings of International Conference on Cluster Computing and Workshops, pp. 1–8.
Shanjiang, T., Bu-Sung, L., Bingsheng, H., 2013. Dynamic slot allocation technique for MapReduce clusters. In: International Conference on Cluster Computing, pp. 1–8.
Shanjiang, T., Bu-Sung, L., Bingsheng, H., 2014. DynamicMR: a dynamic slot allocation optimization framework for MapReduce clusters. IEEE Trans. Cloud Comput. 2 (3), 333–347.
Sharma, B., Wood, T., Das, C., 2013. HybridMR: a hierarchical MapReduce scheduler for hybrid data centers. In: Proceedings of International Conference on Distributed Computing Systems, pp. 102–111.
Sidhanta, S., Golab, W., Mukhopadhyay, S., 2016. OptEx: a deadline-aware cost optimization model for spark. In: IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pp. 193–202.
Singh, N., Agrawal, S., 2015. A review of research on MapReduce scheduling algorithms in Hadoop. In: Proceedings of International Conference on Computing, Communication Automation, pp. 637–642.
Soualhia, M., Khomh, F., Tahar, S., 2015. ATLAS: an adaptive failure-aware scheduler for Hadoop. In: Proceedings of International Performance Computing and Communications Conference, pp. 1–8.
Tang, S., Lee, B.S., He, B., 2016. Dynamic job ordering and slot configurations for MapReduce workloads. IEEE Trans. Serv. Comput. 9 (1), 4–17.
Tang, Z., Liu, M., Ammar, A., Li, K., Li, K., 2016. An optimized MapReduce workflow scheduling algorithm for heterogeneous computing. J. Supercomput. 72 (6), 2059–2079.
Tseng-Yi, C., Hsin-Wen, W., Ming-Feng, W., Ying-Jie, C., Tsan-sheng, H., Wei-Kuan, S., 2013. LaSA: a locality-aware scheduling algorithm for Hadoop-MapReduce resource assignment. In: Proceedings of International Conference on Collaboration Technologies and Systems, pp. 342–346.
Ullah, I., Jihyeon, C., Yonjoong, R., Man, Y., Hee, Y., 2014. Hadoop preemptive deadline constraint scheduler. In: Proceedings of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 201–208.
Wang, W., Zhu, K., Ying, L., Tan, J., Zhang, L., 2013. A throughput optimal algorithm for map task scheduling in MapReduce with data locality. ACM SIGMETRICS Performance Eval. Rev. 40 (4), 33–42.
Wei, Z., Rajasekaran, S., Wood, T., Mingfa, Z., 2014. MIMP: deadline and interference aware scheduling of Hadoop virtual machines. In: Proceedings of International Symposium on Cluster, Cloud and Grid Computing, pp. 394–403.
Wen, Y.-F., 2016. Energy-aware dynamical hosts and tasks assignment for cloud computing. J. Syst. Softw. 115 (C), 144–156.
Wohlin, C., 2014. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of International Conference on Evaluation and Assessment in Software Engineering, pp. 38:1–38:10.
Wolf, J., Rajan, D., Hildrum, K., Khandekar, R., Kumar, V., Parekh, S., Wu, K., balmin, A., 2010. FLEX: A slot allocation scheduling optimizer for MapReduce workloads. In: Proceedings of International Conference on Middleware, pp. 1–20.
Xiaohong, Z., Yuhong, F., Shengzhong, F., Jianping, F., Zhong, M., 2011. An effective data locality aware task scheduling method for MapReduce framework in heterogeneous environments. In: Proceedings of International Conference on Cloud and Service Computing, pp. 235–242.
Xiaohong, Z., Zhiyong, Z., Shengzhong, F., Bibo, T., Jianping, F., 2011. Improving data locality of MapReduce by scheduling in homogeneous computing environments. In: Proceedings of International Symposium on Parallel and Distributed Processing with Applications, pp. 120–126.
Xu, J., Chen, Z., Tang, J., Su, S., 2014. T-Storm: traffic-aware online scheduling in storm. In: Proceedings of IEEE International Conference on Distributed Computing Systems, pp. 535–544.
Xu, J., Chen, Z., Tang, J., Su, S., 2014. T-Storm: traffic-aware online scheduling in storm. In: Proceedings of IEEE International Conference on Distributed Computing Systems, pp. 535–544.
Xu, Z., Xiaoshe, D., Haijun, C., Yuanquan, F., Huo, Z., 2012. A parameter dynamic–tuning scheduling algorithm based on history in heterogeneous environments. In: Proceedings of ChinaGrid Annual Conference, pp. 49–56.
Xue, R., Gao, S., Ao, L., Guan, Z., 2015. BOLAS: bipartite-graph oriented locality-aware scheduling for MapReduce tasks. In: Proceedings of International Symposium on Parallel and Distributed Computing, pp. 37–45.
Yao, Y., Wang, J., Sheng, B., Lin, J., Mi, N., 2014. HaSTE: Hadoop YARN scheduling based on task-dependency and resource-demand. In: Proceedings of IEEE International Conference on Cloud Computing, pp. 184–191.
Yao, Y., Wang, J., Sheng, B., Tan, C., Mi, N., 2015. Self-adjusting slot configurations for homogeneous and heterogeneous Hadoop clusters. IEEE Trans. Cloud Comput. PP (99), 1–14.
Yildiz, O., Ibrahim, S., Antoniu, G., 2017. Enabling fast failure recovery in shared Hadoop clusters: towards failure-aware scheduling. Future Gener. Comput. Syst. 74, 208–219.
Yildiz, O., Ibrahim, S., Phuong, T.A., Antoniu, G., 2015. Chronos: failure-aware scheduling in shared Hadoop clusters. In: Proceedings of IEEE International Conference on Big Data, pp. 313–318.
Yin, L., Chuang, L., Fengyuan, R., Yifeng, G., 2013. H-PFSP: efficient hybrid parallel PFSP protected scheduling for MapReduce system. In: Proceedings of International Conference on Trust, Security and Privacy in Computing and Communications, pp. 1099–1106.
Yuan, Z., Wang, J., 2013. Research of scheduling strategy based on fault tolerance in Hadoop platform. In: Geo-Informatics in Resource Management and Sustainable Ecosystem. Springer, pp. 509–517.
Zaharia, M., Borthakur, D., Sen Sarma, J., Elmeleegy, K., Shenker, S., Stoica, I., 2009. Job scheduling for multi-user MapReduce clusters. Technical Report. EECS Department, University of California, Berkeley, USA.
Zaharia, M., Borthakur, D., Sen Sarma, J., Elmeleegy, K., Shenker, S., Stoica, I., 2010. Delay scheduling: a simple technique for achieving locality and fairness in cluster scheduling. In: Proceedings of European Conference on Computer Systems, pp. 265–278.
Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I., 2012. Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of USENIX Conference on Networked Systems Design and Implementation, pp. 1–14.
Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I., 2010. Spark: cluster computing with working sets. In: Proceedings of USENIX Conference on Hot Topics in Cloud Computing, pp. 1–7.
Zaharia, M., Konwinski, A., Joseph, A., Katz, R., Stoica, I., 2008. Improving MapReduce performance in heterogeneous environments. In: Proceedings of USENIX Conference on Operating Systems Design and Implementation, pp. 29–42.
Zhang, Q., Zhani, M., Yang, Y., Boutaba, R., Wong, B., 2015. PRISM: fine-grained resource-aware scheduling for MapReduce. IEEE Trans. Cloud Comput. 3 (2), 182–194.
Zheng, S., Liu, Y., He, T., Shanshan, L., Liao, X., 2014. SkewControl: gini out of the bottle. In: Proceedings of IEEE International Parallel Distributed Processing Symposium Workshops, pp. 1572–1580.

Mbarka Soualhia holds an M.Sc. degree in Engineering concentration Information Technology from École de Technologie Superieure (ÉTS), Canada and a bachelor degree in Computer Sciences from École Nationale Superieure d'Ingenieurs de Tunis (ÉNSIT), Tunisia. She is currently a Ph. D. candidate at Concordia University, Canada and she is working as research assistant under the supervision of Prof. Sofiène Tahar and Prof. Foutse Khomh. Her research focuses on designing adaptive software components and software architecture to process intensive data applications in distributed systems and their verification using formal methods such as Theorem Proving and Model Checking.

Foutse Khomh is an associate professor at Polytechnique Montréal, where he heads the SWAT Lab on software analytics and cloud engineering research (http://swat.polymtl.ca/). He received a Ph.D in Software Engineering from the University of Montreal in 2010, with the Award of Excellence. His research interests include software maintenance and evolution, cloud engineering, service-centric software engineering, empirical software engineering, and software analytic. He has published several papers in international conferences and journals, including ICSM(E), ASE, ISSRE, SANER, ICWS, HPCC, IPCCC, JSS, JSEP, and EMSE. His work has received three Best Paper Awards and many nominations. He has served on the program committees of several international conferences including ICSM(E), SANER, MSR, ICPC, SCAM, ESEM and has reviewed for top international journals such as SQJ, EMSE, TSE and TOSEM. He is on the Review Board of EMSE. He is program chair for Satellite Events at SANER 2015, program co-chair of SCAM 2015 and ICSME 2018, and general chair of ICPC 2018. He is one of the organizers of the RELENG workshop series (http://releng.polymtl.ca) and has been guest editor for special issues in the IEEE Software magazine and JSEP.

Sofiène Tahar received the Diploma degree in computer engineering from the University of Darmstadt, Germany, in 1990, and the Ph.D. degree with distinction in computer science from the University of Karlsruhe, Germany, in 1994. Currently, he is a professor and the research chair in formal verification of system-on-chip at the Department of Electrical and Computer Engineering, Concordia University. His research interests are in the areas of formal hardware verification, system-on-chip verification, analog and mixed signal circuits verification, and probabilistic, statistical and reliability analysis of systems. Dr. Tahar, a professional engineer in the Province of Quebec, is the founder and director of the Hardware Verification Group at Concordia University. He is a senior member of ACM and a senior member of IEEE.