arXiv:1312.4722v1 [cs.DC] 17 Dec 2013

Big Data Computing and Clouds: Challenges, Solutions, and Future Directions

Marcos D. Assunção[a], Rodrigo N. Calheiros[b], Silvia Bianchi[a], Marco A. S. Netto[a], Rajkumar Buyya[b,*]

[a] *IBM Research, Brazil*
[b] *The University of Melbourne, Australia*

Abstract

This paper discusses approaches and environments for carrying out analytics on Clouds for Big Data applications. It revolves around four important areas of analytics and Big Data, namely (i) data management and supporting architectures; (ii) model development and scoring; (iii) visualisation and user interaction; and (iv) business models. Through a detailed survey, we identify possible gaps in technology and provide recommendations for the research community on future directions on Cloud-supported Big Data computing and analytics solutions.

Keywords:
Big Data, Cloud Computing, Analytics, Data Management

1. Introduction

Society is becoming increasingly more instrumented and as a result, organisations are producing and storing vast amounts of data. Managing and gaining insights from the produced data is a challenge and key to competitive advantage. Analytics solutions that mine structured and unstructured data are important as they can help organisations gain insights not only from their privately acquired data, but also from large amounts of data publicly available on the Web [1]. The ability to cross relate private information on consumer preferences and products with information from tweets, blogs,

*Corresponding author: rbuyya@unimelb.edu.au

product evaluations, and data from social networks opens a wide range of possibilities for organisations to understand the needs of their customers, predict their wants and demands, and optimise the use of resources. This paradigm is being popularly termed as Big Data.

Despite the popularity on analytics and Big Data, putting them into practice is still a complex and time consuming endeavour. As highlighted by Yu [2], Big Data offers substantial value to organisations willing to adopt it, but at the same time poses a considerable number of challenges for the realisation of such added value. An organisation willing to use analytics technology frequently acquires expensive software licenses; employs large computing infrastructure; and pays for consulting hours of analysts who work with the organisation to better understand its business, organise its data, and integrate it for analytics [3]. This joint effort of organisation and analysts often aims to help the organisation understand its customers' needs, behaviours, and future demands for new products or marketing strategies. Such effort, however, is generally costly and often lacks flexibility. Nevertheless, research and application of Big Data are being extensively explored by governments, as evidenced by initiatives from the governments from USA [4] and UK [5]; by academics, such as the bigdata@csail initiative from MIT [6]; and by companies such as Intel [7].

Cloud computing has been revolutionising the IT industry by adding flexibility to the way IT resources are consumed, enabling organisations to pay only for the resources and services they use. In an effort to reduce IT capital and operational expenditures, organisations of all sizes are using Clouds to provide the resources required to run their applications. Clouds vary significantly in their specific technologies and implementation, but often provide infrastructure, platform, and software resources as services [8, 9].

The most often claimed benefits of Clouds include offering resources in a pay-as-you-go fashion, improved availability and elasticity, and cost reduction. Clouds can prevent organisations from spending money maintaining peak-provisioned IT infrastructure they are unlikely to use most of the time. While at first glance the value proposition of Clouds as a platform to carry out analytics is strong, there are many challenges that need to be overcome to make Clouds an ideal platform for scalable analytics.

In this article we survey approaches and environments on areas that are key to Big Data analytics capabilities and discuss how they help building analytics solutions for Clouds. We focus on the most important technical issues on enabling Cloud analytics, but also highlight some of the non-technical

challenges faced by organisations that want to provide analytics on the Cloud.

2. Background

Organisations are increasingly generating large volumes of data as result of instrumented business processes, monitoring of user activity [10, 11], web site tracking, sensors, finance, accounting, among other reasons. With the advent of social network websites, users create records of their lives by daily posting details of activities they perform, events they attend, places they visit, pictures they take, and things they enjoy and want. This data deluge is often referred to as Big Data [12, 13, 14]; a term that conveys the challenges it poses on existing infrastructure in respect to storage, management, interoperability, governance, and analysis of the data.

In today's competitive market, being able to explore data to understand customer behaviour, segment customer base, offer customised services, and gain insights from data provided by multiple sources is key to competitive advantage. Whilst decision makers would like to base their decisions and actions on insights gained from this data [15], making sense of data, extracting non obvious patterns, and using these patterns to predict future behaviour are not new topics. Knowledge Discovery in Data (KDD) [16] aims to extract non obvious information using careful and detailed analysis and interpretation. Data mining [17, 18], more specifically, aims to discover previously unknown interrelations among apparently unrelated attributes of datasets by applying methods from several areas including machine learning, database systems, and statistics. Analytics comprises techniques of KDD, data mining, text mining, statistical and quantitative analysis, explanatory and predictive models, and advanced and interactive visualisation to drive decisions and actions [19, 15, 20].

Figure 1 depicts the common phases of a traditional analytics workflow for Big Data. Data from various sources, including databases, streams, marts, and data warehouses, are used to build models. The large volume and different types of the data can demand pre-processing tasks for integrating the data, cleaning it, and filtering it. The prepared data is used to train a model and to estimate its parameters. Once the model is estimated, it should be validated before its consumption. Normally this phase requires the use of the original input data and specific methods to validate the created model. Finally, the model is consumed and applied to data as it arrives. This phase, called model scoring, is used to generate predictions, prescriptions, and rec-

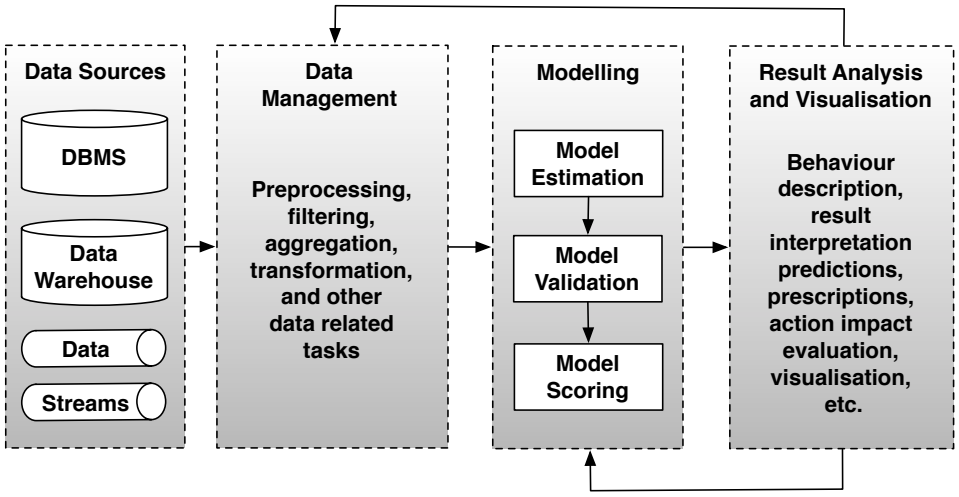

Figure 1: Overview of the analytics workflow for Big Data.

ommendations. The results are interpreted and evaluated, used to generate new models or calibrate existing ones, or are integrated to pre-processed data.

Analytics solutions can be classified as descriptive, predictive, or prescriptive as illustrated in Figure 2. Descriptive analytics uses historical data to identify patterns and create management reports; it is concerned with modelling past behaviour. Predictive analytics attempts to predict the future by analysing current and historical data. Prescriptive solutions assist analysts in decisions by determining actions and assessing their impact regarding business objectives, requirements, and constraints.

Despite the hype about analytics, using analytics is still a labour intensive endeavour, often requiring expensive software and several consulting hours, or even days, to develop and tailor a solution to an organisation's specific business needs [3]. Such solutions are often developed and hosted on the customer's premises, are generally complex, and their operations can take hours do execute. Cloud computing provides an interesting model for analytics, where solutions can be hosted in the Cloud and consumed by customers in a pay-as-you-go fashion. For this delivery model to become reality, however, several technical issues must be addressed, such as data management, tuning of models, privacy, data quality, and data currency.

This work highlights technical issues and surveys existing work on solu-

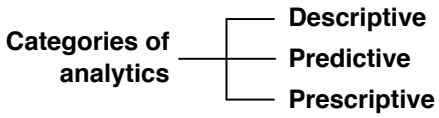

Figure 2: Categories of analytics.

tions to provide analytics capabilities for Big Data on the Cloud. Considering the traditional analytics workflow presented in Figure 1, we focus on key issues in the phases of an analytics solution. With Big Data it is evident that many of the challenges of Cloud analytics concern data management, integration, and processing. Previous work has focused on issues such as data formats, data representation, storage, access, privacy, and data quality. Section 3 presents existing work addressing these challenges on Cloud environments. In Section 4, we elaborate on existing models to provide and evaluate data models on the Cloud. Section 5 describes solutions for data visualisation and customer interaction with analytics solutions provided by a Cloud. We also highlight some of the business challenges posed by this delivery model when we discuss service structures, service level agreements, and business models. Security is certainly a key challenge for hosting analytics solutions on public Clouds. We consider, however, that security is an extensive topic and would therefore deserve a study of its own. Therefore, security and evaluation of data correctness [21] are left out of scope of this survey.

3. Data Management

One of the most time-consuming and labour-intensive tasks of analytics is preparation of data for analysis; a problem often exacerbated by Big Data as it stretches existing infrastructure to its limits. Performing analytics on large volumes of data requires efficient methods to store, filter, transform, and retrieve the data. Some of the challenges of deploying data management solutions on Cloud environments have been known for some time [22, 23, 24], and solutions to perform analytics on the Cloud face similar challenges. Cloud analytics solutions need to consider the multiple Cloud deployment models adopted by enterprises, where Clouds can be for instance:

- *Private*: deployed on a private network, managed by the organisation itself or by a third party. A private Cloud is suitable for businesses

that require the highest level of control of security and data privacy. In such conditions, this type of Cloud infrastructure can be used to share the services and data more efficiently across the different departments of a large enterprise.

- *Public*: deployed off-site over the Internet and available to the general public. Public Cloud offers high efficiency and shared resources with low cost. The analytics services and data management are handled by the provider and the quality of service (*e.g.* privacy, security, and availability) is specified in a contract. Organisations can leverage these Clouds to carry out analytics with a reduced cost or share insights of public analytics results.

- *Hybrid*: combines both Clouds where additional resources from a public Cloud can be provided as needed to a private Cloud. Customers can develop and deploy analytics applications using a private environment, thus reaping benefits from elasticity and higher degree of security than using only a public Cloud.

Considering the Cloud deployments, the following scenarios are generally envisioned regarding the availability of data and analytics models [25]: (i) data and models are private; (ii) data is public, models are private; (iii) data and models are public; and (iv) data is private, models are public. Jensen *et al.* [26] advocate on deployment models for Cloud analytics solutions that vary from solutions using privately hosted software and infrastructure, to private analytics hosted on a third party infrastructure, to public model where the solutions are hosted on a public Cloud.

Different from traditional Cloud services, analytics deals with high-level capabilities that often demand very specialised resources such as data and domain experts' analysis skills. For this reason, we advocate that under certain business models—specially those where data and models reside on the provider's premises—not only ordinary Cloud services, but also the skills of data experts need to be managed. To achieve economies of scale and elasticity, Cloud-enabled Big Data analytics needs to explore means to allocate and utilise these specialised resources in a proper manner. The rest of this section discusses existing solutions on data management irrespective of where data experts are physically located, focusing on storage and retrieval of data for analytics; data diversity, velocity and integration; and resource scheduling for data processing tasks.

3.1. Data Variety and Velocity

Big Data is characterised by what is often referred to as a multi-V model, as depicted in Figure 3, where variety, velocity, and volume [27, 28] are the items most commonly mentioned. Variety represents the data types, velocity refers to the rate at which the data is produced and processed, and volume defines the amount of data. Veracity refers to how much the data can be trusted given the reliability of its source [2].

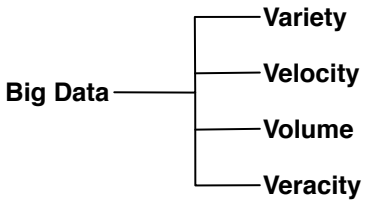

Figure 3: The '4 Vs' of Big Data.

Regarding Variety, it can be observed that over the years, a lot of data has been made publicly available for scientific and business uses. Examples include repositories with government statistics; historical weather information and forecasts; DNA sequencing; information on traffic conditions in large metropolitan areas; product reviews and comments; demographics [29]; comments, pictures, and videos posted on social network websites; and data collected by a multitude of sensors measuring various environmental conditions such as temperature, air humidity, air quality, and precipitation. An example illustrating the need for such a variety within a single analytics application is the Eco-Intelligence [30] platform. Eco-Intelligence was designed to analyse large amounts of data to support city planning and promote more sustainable development. The platform aims to efficiently discover and process data from several sources, including sensors, news, Web sites, television and radio, and exploit information to help urban stakeholders cope with the highly dynamics of urban development. In a related scenario, the Mobile Data Challenge (MDC) was created aimed at generating innovations on smartphone-based research, and to enable community evaluation of mobile data analysis methodologies [31]. Data from around 200 users of mobile phones was collected over a year as part of the Lausanne Data Collection Campaign. Another related area benefiting from analytics is Massively Multiplayer Online Games (MMOGs). CAMEO [32] is an architecture for continuous analytics for MMOGs that uses Cloud resources for analysis tasks. The architecture

provides mechanisms for data collection and continuous analytics on several factors such as understanding the needs of the game community.

Data is also often available for sale in the format of research and technical reports, market segment and financial analyses, among other means. This data can be used by various applications, for instance, to improve the living conditions in large cities, to provide better quality services, to optimise the use of natural resources[1], and to prevent or manage response to unplanned events.

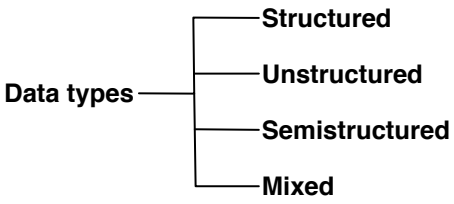

Figure 4: Variety of data.

Handling and analyzing this data poses several challenges because as shown in Figure 4, the data can be of different types. It is argued that a large part of data produced today is either *unstructured* or *semistructured*.

Regarding data Velocity, it is noticed that, to complicate matters further, data can arrive and require processing at different speeds, as illustrated in Figure 5. While for some applications, the arrival and processing of data can be performed in batch, other analytics applications require continuous and real-time analyses, sometimes requiring immediate action upon processing of incoming data streams. For instance, to provide active management for data centres, Wang *et al.* [33] present an architecture that integrates monitoring and analytics. The proposed architecture relies on Distributed Computation Graphs (DCG) that are created to implement the desired analytics functions. The motivating use cases consist in scenarios where information can be collected from monitored equipments and services, and once a potential problem is identified, the system can instantiate DCGs to collect further information for analytics.

Increasingly often, data arriving via streams needs to be analysed and compared against historical information. Different data sources may use

[1]http://www.sense-t.org.au/about/the-big-picture

their own formats, which makes it difficult to *integrate* data from multiple sources in an analytics solution. As highlighted in existing work [34], *standard formats and interfaces* are crucial so that solution providers can benefit from economies of scale derived from data integration capabilities that address the needs of a wide range of customers.

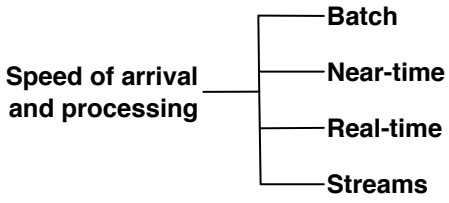

Figure 5: Velocity of data.

The rest of our discussion on data management for Cloud analytics surrounds these two Vs of Big Data, namely Variety and Velocity. We survey solutions on how this diverse data is stored, how it can be integrated and how it is often processed. The discussion on visualisation also explores Velocity by highlighting factors such as interactivity and batch based visualisation. Although the other Vs of Big Data are important, we consider that some of them, as discussed earlier, deserve a study of their own, such as data Veracity. Regarding Volume, in addition to being a little subjective, it is highly dependent on the scalability of existing hardware infrastructure, which improves quickly and can render a survey obsolete very rapidly.

3.2. Data Storage

Several solutions were proposed to store and retrieve large amounts of data demanded by Big Data, some of which are currently used in Clouds. Internet-scale file systems such as the Google File System (GFS) [35] attempt to provide the robustness, scalability, and reliability that certain Internet services need. Other solutions provide object-store capabilities where files can be replicated across multiple geographical sites to improve redundancy, scalability, and data availability. Examples include Amazon Simple Storage Service (S3)[2], Nirvanix Cloud Storage[3], OpenStack Swift[4] and Windows

[2]http://aws.amazon.com/s3/
[3]http://www.nirvanix.com
[4]http://swift.openstack.org

Azure Binary Large Object (Blob) storage[5]. Although these solutions provide the scalability and redundancy that many Cloud applications require, they sometimes do not meet the concurrency and performance needs of certain analytics applications.

MapReduce presents an interesting model where data is placed more closely to where it is processed. Hadoop, an open source MapReduce implementation, allows for the creation of clusters that use the Hadoop Distributed File System (HDFS) to partition and replicate datasets to nodes where they are more likely to be consumed by mappers. In addition to handling embarrassingly parallel applications by exploiting concurrency of large numbers of nodes, HDFS minimises the impact of failures by replicating datasets to a configurable number of nodes. It has been used by Thusoo *et al.* [36] to develop an analytics platform to process Facebook's large data sets. The platform uses Scribe to aggregate logs from Web servers and then exports them to HDFS files and uses a Hive-Hadoop cluster to execute analytics jobs. The platform includes replication and compression techniques and columnar compression of Hive[6] to store the large amounts of data.

Among the drawbacks of Cloud storage techniques and MapReduce implementations, there is the fact that they require the customer to learn a new set of APIs to build analytics solutions for the Cloud. To minimise this hurdle, previous work has also investigated POSIX-like file systems for data analytics. As an example, Ananthanarayana *et al.* [37] adapted POSIX-based cluster file systems to be used as data storage for Cloud analytics applications. By using the concept of meta-blocks, they demonstrated that IBM's General Parallel File System (GPFS) [38] can match the read performance of HDFS. A meta-block is a consecutive set of data blocks that are allocated in the same disk, thus guaranteeing contiguity. The proposed approach explores the trade-off between different block sizes, where meta-blocks minimise seek overhead in MapReduce applications, whereas small blocks reduce pre-fetch overhead and improves cache management for ordinary applications. Tantisiriroj *et al.* [39] compared the Parallel Virtual File System (PVFS) [40] against HDFS, where they observed that PVFS did not present significant improvement in completion time and throughput compared to HDFS.

Although a large part of the data produced nowadays is unstructured,

[5]http://www.windowsazure.com/en-US/services/data-management/

[6]http://hive.apache.org

relational databases have been the choice most organisations have made to store data about their customers, sales, and products, among other things. As data managed by traditional DBMS ages, it is moved to data warehouses for analysis and for sporadic retrieval. Models such as MapReduce are generally not the most appropriate to analyze such relational data. Attempts have been made to provide hybrid solutions that incorporate MapReduce to perform some of the queries and data processing required by DBMS's [22]. Cohen *et al.* [41] provide a parallel database design for analytics that supports SQL and MapReduce scripting on top of a DBMS to integrate multiple data sources. A few providers of analytics and data mining solutions, by exploring models such as MapReduce, are migrating some of the processing tasks closer to where the data is stored, thus trying to minimise surpluses of siloed data preparation, storage, and processing [42]. Data processing and analytics capabilities are moving towards Enterprise Data Warehouses (EDWs), or are being deployed in data hubs [26] to facilitate reuse across various data sets.

In respect to EDW, some Cloud providers offer solutions that promise to scale to one petabyte of data or more. Amazon Redshift [43], for instance, offers columnar storage and data compression and aims to deliver high query performance by exploring a series of features, including a massively parallel processing architecture using high performance hardware, mesh networks, locally attached storage, and zone maps to reduce the I/O required by queries. Amazon Data Pipeline [44] allows a customer to move data across different Amazon Web Services, such as Elastic MapReduce (EMR) [45] and DynamoDB [46], and hence compose the required analytics capabilities.

Another distinctive trend in Cloud computing is the increasing use of NoSQL databases as the preferred method for storing and retrieving information. NoSQL adopts a non-relational model for data storage. Leavitt argues that non-relational models are available for more than 50 years in forms such as object-oriented, hierarchical, and graph databases, but recently this paradigm started to attract more attention with models such as key-store, column-oriented, and document-based stores [47]. The causes for such raise in interest, according to Levitt, are better performance, capacity of handling unstructured data, and suitability for distributed environments [47].

Han *et al.* [48] presented a survey of NoSQL databases with emphasis on their advantages and limitations for Cloud computing. The survey classifies NoSQL systems according to their capacity in addressing different pairs of CAP (consistency, availability, partitioning). The survey also explores the data model that the studied NoSQL systems support.

Hecht and Jablonski [49] compared different NoSQL systems in regard to supported data models, types of query supported, and support for concurrency, consistency, replication, and partitioning. Hecht and Jablonski concluded that there are big differences among the features of different technologies, and there is no single system that would be the most suitable for every need. Therefore, it is important for adopters to understand the requirements of their applications and the capabilities of different systems so the system whose features better match their needs is selected [49].

3.3. Data Integration Solutions

Forrester Research realised a technical report that discusses some of the problems that traditional Business Intelligence (BI) faces [42], highlighting that there is often a surplus of siloed data preparation, storage, and processing. Authors of the report envision some data processing and Big Data analytics capabilities being migrated to the EDW, hence freeing organisations from unnecessary data transfer and replication and the use of disparate data-processing and analysis solutions. Moreover, as discussed earlier, they advocated that analytics solutions will increasingly expose data processing and analysis features via MapReduce and SQL-MR-like interfaces. SAP HANA One [50], as an example, is a in-memory platform hosted by Amazon Web Services that provides real-time analytics for SAP applications. HANA One also offers a SAP data integrator to load data from HDFS and Hive-accessible databases.

EDWs or Cloud based data warehouses, however, create certain issues in respect to data integration and the addition of new data sources. Standard formats and interfaces can be essential to achieve economies of scale and meet the needs of a large number of customers [34]. Some solutions attempt to address some of these issues [51, 29]. Birst [51] provides composite spaces and space inheritance, where a composite space integrates data from one or more parent spaces with additional data added to the composite space. Birst provides a Software as a Service (SaaS) solution that offers analytics functionalities on a subscription model; and appliances with the business analytics infrastructure, hence providing a model that allows a customer to migrate gradually from an on-premise analytics to a scenario with Cloud-provided analytics infrastructure. To improve the market penetration of analytics solutions in emerging markets such as India, Deepak *et al.* [52] propose a multi-flow solution for analytics that can be deployed on the Cloud. The multi-flow approach provides a range of possible analytics operators and

flows to compose analytics solutions; viewed as workflows or instantiations of a multi-flow solution. IVOCA [53] is a tool aimed at Customer Relationship Management (CRM) that ingests both structured and unstructured data and provides data linking, classification, and text mining tools to facilitate analysts' tasks and reduce the time to insight.

Habich *et al.* [54] propose Web services that co-ordinate data Clouds for exchanging massive data sets. The Business Process Execution Language (BPEL) data transition approach is used for data exchange by passing references to data between services to reduce the execution time and guarantee the correct data processing of an analytics process. A generic data Cloud layer is introduced to handle heterogeneous data Clouds, and is responsible for mapping generic operations to each Cloud implementation. DataDirect Cloud [55] also provides generic interfaces by offering JDBC/ODBC drivers for applications to execute SQL queries against different databases stored on a Cloud. Users are not required to deal with different APIs and query languages specific to each Cloud storage solution.

PivotLink's AnalyticsCLOUD [29] handles both structured and unstructured data, providing data integration features. PivotLink also provides DataCLOUD with information about over 350 demographic, hobbies, and interest data fields for 120 million U.S. households. This information can be used by customers to perform brand sentiment analysis [56] and verify how weather affects their product performance.

3.4. Data Processing and Resource Management

MapReduce [57] is one of the most popular programming models to process large amounts of data via embarrassingly parallel applications running on clusters of computers. Hadoop [58] is the most used open source MapReduce implementation, also made available by several Cloud providers [45, 59, 60, 61]. Amazon EMR [45] enables customers to instantiate Hadoop clusters to process large amounts of data using the Amazon Elastic Compute Cloud (EC2) and other Amazon Web Services for data storage and transfer.

Hadoop uses the HDFS file system to partition and replicate data sets across multiple nodes, such that when running a MapReduce application, a mapper is likely to access data that is locally stored on the cluster node where it is executing. Although Hadoop provides a set of APIs that allow developers to implement MapReduce applications, very often a Hadoop workflow is composed of jobs that use high-level query languages such as Hive and Pig Latin, created to facilitate search and specification of processing tasks.

Lee *et al.* [62] present a survey about the features, benefits, and limitations of MapReduce for parallel data analytics. They also discuss extensions proposed for this programming model to overcome some of its limitations.

Hadoop achieves interesting levels of parallelism and fault tolerance, but what is often criticised about it is the time required to load data into HDFS and the lack of reuse of data produced by mappers. MapReduce is a model created to exploit commodity hardware, but when executed on reliable infrastructure, the mechanisms it provides to deal with failures may not be entirely essential. Some of the provided features can be disabled in certain scenarios. Herodotou and Babu [63] present techniques for profiling MapReduce applications, identifying bottlenecks and simulating what-if scenarios. Previous work has also proposed optimisations to handle these shortcomings [64]. Cuzzocrea *et al.* [65] discuss issues concerning analytics over big multidimensional data and the difficulties in building multidimensional structures in HDFS and integrating multiple data sources to Hadoop.

Starfish [66], a data analytics system built atop Hadoop, focuses on improving the performance of clusters throughout the data lifecycle in analytics, without requiring users to understand the available configuration options. Starfish employs techniques at several levels to optimise the execution of MapReduce jobs. It uses dynamic instrumentation to profile jobs and optimises workflows by minimising the impact of data unbalance and by balancing the load of executions. Starfish's Elastisizer automates provisioning decisions using a mix of simulation and model-based estimation to address what-if questions on workload performance.

Lee *et al.* [67] present an approach that allocates resources and schedules jobs considering data analytics workloads, in order to enable consolidation of a cluster workload, reducing the number of machines allocated for processing the workload during periods of small load. The approach uses Hadoop and works with two pools of machines—*core* and *accelerator*—and dynamically adjusts the size of each pool according to the observed load.

Daytona [59], a MapReduce runtime for Windows Azure, leverages the scalable storage services provided by Azure's Cloud infrastructure as the source and destination of data. It uses Cloud features to provide load balancing and fault tolerance. The system relies on a master-slave architecture where the master is responsible for scheduling tasks and the slaves for carrying out map and reduce operations. Section 5 discusses the visualisation features that Daytona provides.

Previous work shows that there is an emerging class of MapReduce ap-

plications that feature small, short, and highly interactive jobs [68, 69]. As highlighted in Section 5, the visualisation community often criticises the lack of interactivity of MapReduce-based analytics solutions. Over the past few years, however, several attempts have been made to tackle this issue [70]. Borthakur *et al.* for instance, describe optimisations implemented in HDFS and HBase[7] to make them more responsive to the realtime requirements of Facebook applications. Chen *et al.* [71] propose energy efficiency improvements to Hadoop by maintaining two distinct pools of resources, namely to interactive and batch jobs.

The eXtreme Analytics Platform (XAP) [72] enables analytics supporting multiple data sources, data types (structured and unstructured), and multiple types of analyses. The target infrastructure of the architecture is a cluster running a distributed file system. A modified version of Hadoop, deployed in the cluster, contains an application scheduler (FLEX) able to better utilise the available resources than the default Hadoop scheduler. The analytics jobs are created via a high-level language script, called Jaql, that converts the high-level descriptive input into an analytics MapReduce workflow that is executed in the target infrastructure.

Previous work has also considered other models for performing analytics, such as scientific workflows and Online Analytical Processing (OLAP). Rahman *et al.* [73] propose a hybrid heuristic for scheduling data analytics workflows on heterogeneous Cloud environments; a heuristic that optimises cost of workflow execution and satisfies users requirements, such as budget, deadline, and data placement. In the field of simulation-enabled analytics, Li *et al.* [74] developed an analytical application, modelled as a Direct Acyclic Graph (DAG), for predicting the spread of dengue fever outbreaks in Singapore. The analytics workflow receives data from multiple sources, including current and past data about climate and weather from meteorological agencies and historical information about dengue outbreaks in the country. This data, with user-supplied input about the origin of the infection, is used to generate a map of the spread of the disease in the country in a day-by-day basis. A hybrid Cloud is used to speed up the application execution. Other characteristics of the application are security features and cost-effective exploration of Cloud resources: the system keeps the utilisation of public Cloud resources to a minimum to enable the analytics to complete in the specified

[7] http://hbase.apache.org

time and budget. A public Cloud has also been used in a similar scenario to simulate the impact of public transport disruptions on urban mobility [75]. CloudComet [76] was used to perform online risk analyses; its programming layer provides a framework for application development and management that supports a range of paradigms, including master-worker, bag-of-tasks, and MapReduce.

The support of OLAP for Google App Engine (GAE) [77] was evaluated [78], highlighting limitations and evaluating their impact on cost and performance of applications. A hybrid approach to perform OLAP using GAE and AppScale [79] was provided, using two methods for data synchronisation, namely *bulk data transfer* and *incremental data transfer*. Moreover, Jung *et al.* [80] propose optimisations for scheduling and processing of Big Data analysis on federated Clouds.

Chang *et al.* [81] examined different data analytics workloads, where results show significant diversity of resource usage (CPU, I/O and, network). They recommend the use of transformation mechanisms such as indexing, compression, and approximation to provide a balanced system and improve efficiency of data analysis.

The Cloud can also be used to extend the capabilities of analyses initially started on the customer's premises. CometCloud, for example, is an autonomic computing engine that supports Cloud bursts that has been used to provide the programming and runtime infrastructure to scale out and in certain on-line risk analyses [76]. CloudComet and commercial technologies such as Aneka [82] can utilise both private resources and resources from a public Cloud provider to handle peaks in the demands of online risk analytics.

Some analytics applications including stock quotes and weather prediction have stringent time constraints, usually falling in the near-time and streams categories described earlier. Request processing time is important to deliver results in a timely fashion. Chen *et al.* [83] investigate Continuous analytics as a Service (CaaaS) that blends stream processing and relational data techniques to extend the DBMS model and enable real-time continuous analytics service provisioning. The dynamic stream processing and static data management for data intensive analytics are unified by providing an SQL-like interface to access both static and stream data. The proposed cycle-based query model and transaction model allow SQL queries to run and to commit per cycle while analysing stream data per chunk. The analysis results are made visible to clients while a continued query for results generation is still running. Existing work on stream and near-time processing attempt to

leverage strategies to predict user or service behaviour [84]. In this way, an analytics service can pre-fetch data to anticipate a user's behaviour, hence selecting the appropriate applications and methods before the user's request arrives.

3.5. Challenges in Big Data Management

In this section, we discussed current research targeting the issue of Big Data management for analytics. There are still, however, many open challenges in this topic. The list below is not exhaustive, and as more research in this field is conducted, more challenging issues will arise.

Data variety: How to handle an always increasing volume of data? Especially when the data is unstructured, how to quickly extract meaningful content out of it? How to aggregate and correlate streaming data from multiple sources?

Data storage: How to efficiently recognise and store important information extracted from unstructured data? How to store large volumes of information in a way it can be timely retrieved? Are current file systems optimised for the volume and variety demanded by analytics applications? If not, what new capabilities are needed? How to store information in a way that it can be easily migrated/ported between data centers/Cloud providers?

Data integration: New protocols and interfaces for integration of data that are able to manage data of different nature (structured, unstructured, semi-structured) and sources.

Data Processing and Resource Management: New programming models optimised for streaming and/or multidimensional data; new backend engines that manage optimised file systems; engines able to combine applications from multiple programming models (*e.g.* MapReduce, workflows, and bag of tasks) on a single solution/abstraction. How to optimise resource usage and energy consumption when executing the analytics application?

4. Model Building and Scoring

The data storage and Data as a Service (DaaS) capabilities provided by Clouds are important, but for analytics, it is equally relevant to use the

Table 1: Summary of works on model building and scoring.

Work	Goal	Service model	Deployment Model
Grazzelli *et al.* [88]	Predictive analytics (scoring)	IaaS	Public
Zementis [89]	Data Analysis and Model Building	SaaS	Public or Private
Google Prediction API [90]	Model Building	SaaS	Public
Apache Mahout [86]	Data Analysis and Model Building	IaaS	Any
Hazy [91]	Model Building	IaaS	Any

data to build models that can be utilised for forecasts and prescriptions. Moreover, as models are built based on the available data, they need to be tested against new data in order to evaluate their ability to forecast future behaviour. Existing work has discussed means to offload such activities—termed here as model building and scoring—to Cloud providers and ways to parallelise certain machine learning algorithms [85, 86, 87]. This section describes work on the topic. Table 1 summarises the analysed work, its goals, and target infrastructures.

Grazzelli *et al.* [88] use Amazon EC2 as a hosting platform for the Zementis' ADAPA model [89] scoring engine. Predictive models, expressed in Predictive Model Markup Language (PMML) [92], are deployed in the Cloud and exposed via Web Services interfaces. Users can access the models with Web browser technologies to compose their data mining solutions. Existing work also advocates the use of PMML as a language to exchange information about predictive models [93].

Zementis [89] also provides technologies for data analysis and model building that can run either on a customer's premises or be allocated as SaaS using Infrastructure as a Service (IaaS) provided by solutions such as Amazon EC2 and IBM SmartCloud Enterprise [94].

Google Prediction API [90] allows users to create machine learning models to predict numeric values for a new item based on values of previously submitted training data or predict a category that best describes an item. The prediction API allows users to submit training data as comma separated files following certain conventions, create models, share their models or use models that others shared. With the Google Prediction API, users

can develop applications to perform analytics tasks such as sentiment analysis [56], purchase prediction, provide recommendations, analyse churn, and detect spam. The Apache Mahout project [86] aims to provide tools to build scalable machine learning libraries on top of Hadoop using the MapReduce paradigm. The provided libraries can be deployed on a Cloud and be explored to build solutions that require clustering, recommendation mining, document categorisation, among others.

By trying to ease the complexity in building trained systems such as IBM's Watson, Apple's Siri and Google Knowledge Graph, the Hazy project [91] focuses on identifying and validating two categories of abstractions in building trained systems, namely *programming abstractions* and *infrastructure abstractions*. It is argued that, by providing such abstractions, it would be easier for one to assemble existing solutions and build trained systems. To achieve a small and compoundable programming interface, Hazy employs a data model that combines the relational data model and a probabilistic rule-based language. For infrastructure abstraction, Hazy leverages the observation that many statistical analysis algorithms behave as a user-defined aggregate in a Relational Database Management System (RDBMS). Hazy then explores features of the underlying infrastructure to improve the performance on these aggregates.

4.1. Open Challenges

The key challenge in the area of Model Building and Scoring is the discovery of techniques that are able to explore the rapid elasticity and large scale of Cloud systems. Given that the amount of data available for Big Data analytics is increasing, timely processing of such data for building and scoring would give a relevant advantage for businesses able to explore such a capability.

In the same direction, standards and interfaces for these activities are also required, as they would help to disseminate "prediction and analytics as services" providers that would compete for customers. If the use of such services does not incur vendor lock in (via utilisation of standards APIs and formats), customers can choose the service provider only based on cost and performance of services, enabling the emergence of a new competitive market.

5. Visualisation and User Interaction

With the increasing amounts of data with which analyses need to cope, good visualisation tools are crucial. These tools should consider the quality of data and presentation to facilitate navigation [95]. The type of visualisation may need to be selected according to the amount of data to be displayed, to improve both displaying and performance. Visualisation can assist in the three major types of analytics: descriptive, predictive, and prescriptive. Many visualisation tools do not describe advanced aspects of analytics, but there has been an effort to explore visualisation to help on predictive and prescriptive analytics, using for instance sophisticated reports and storytelling [96]. A key aspect to be considered on visualisation and user interaction in the Cloud is that network is still a bottleneck in several scenarios [97]. Users ideally would like to visualise data processed in the Cloud having the same experience and feel as though data were processed locally. Some solutions have been tackling this requirement.

For example, as Fisher *et al.* [34] point out, many Cloud platforms available to process data analytics tasks still resemble the *batch-job* model used in the early times of the computing era. Users typically submit their jobs and wait until the execution is complete to download and analyse sample results to validate full runs. As this back and forth of data is not well supported by the Cloud, the authors issue a call to arms for both research and development of better interactive interfaces for Big Data analytics where users iteratively pose queries and see rapid responses. Fisher *et al.* introduce *sampleAction* [98] to explore whether interactive techniques acting over only incremental samples can be considered as sufficiently trustworthy by analysts to make closer to real time decisions about their queries. Interviews with three teams of analysts suggest that representations of incremental query results were robust enough so that analysts were prepared either to abandon a query, refine it, or formulate new queries. King [18] also highlights the importance of making the analytics process iterative, with multiple checkpoints for assessment and adjustment.

In this line, existing work aims to explore the batch-job model provided by solutions including MapReduce as a backend to features provided in interfaces with which users are more familiar. Trying to leverage the popularity of spreadsheets as a tool to manipulate data and perform analysis, Barga *et al.* proposed an Excel ribbon connected to Daytona [59], a Cloud service for data storage and analytics. Users manipulate data sets on Excel and plugins

use Microsoft's Azure infrastructure [99] to run MapReduce applications. In addition, as described earlier, several improvements have been proposed to MapReduce frameworks to handle interactive applications [71, 70, 100]. However, most of these solutions are not yet made available for general use in the Cloud.

Several projects attempt to provide a range of visualisation methods from which users can select a set that suits their requirements. ManyEyes [101] from IBM allows users to upload their data, select a visualisation method—varying from basic to advanced—and publish their results. Users may also navigate through existing visualisations and discuss their findings and experience with peers.

Existing work also provides means for users to aggregate data from multiple sources and employ various visualisation models, including dashboards, widgets, line and bar charts, demographics, among other models [29, 102, 103, 104, 105]. Some of these features can be leveraged to perform several tasks, including create reports; track what sections of a site are performing well and what kind of content can create better user experience; how information sharing on a social network impacts the web site usage; track mobile usage [11, 10]; and evaluate the impact of advertising campaigns.

Choo and Park [106] argue that the reason why Big Data visualisation is not real time is the computational complexity of the analytics operations. In this direction, authors discuss strategies to reduce computational complexity of data analytics operations by, for instance, decreasing precision of calculations.

Apart from software optimisation, hardware specific for visualisation is becoming key for Big Data analytics. For example, Reda *et al.* [107] discuss that, although existing tools are able to provide data belonging to a range of classes, their dimensionality and volume exceed the capacity of visualisation provided by standard displays. This requires the utilisation of large-scale visualisation environments, such as CyberCommons and CAVE2, which are composed of a large display wall with resolution three orders of magnitude higher than that achieved by commercial displays [107]. Remote visualisation systems, such as Nautilus from XSEDE (Extreme Science and Engineering Discovery Environment—the new NSF TeraGrid project replacement), are becoming more common to supply high demand for memory and graphical processors to assist very large data visualisation [108].

Besides visualisation of raw data, summarised content in form of reports are essential to perform predictive and prescriptive analytics. Several solu-

tions have explored report generation and visualisation. For instance, SAP Crystal Solutions [109] provides BI functionalities via which customers can explore available data to build reports with interactive charts, what-if scenarios, and dashboards. The produced reports can be visualised on the Web, e-mail, Microsoft Office, or be embedded into enterprise applications. Another example on report visualisation is Cloud9 Analytics [110], which aims to automate reports and dashboards, based on data from CRM and other systems. It provides features for sales reports, sales analytics, and sales forecasts and pipeline management. By exploring history data and using the notion of risk, it offers customers clues on which projects they should invest their resources and what projects or products require immediate action. Other companies also offer solutions that provide sales forecasts, change analytics, and customised reports [111, 51]. Salesforce [112] supports customisable dashboards through collaborative analytics. The platform allows authorised users to share their charts and information with other users. Another trend on visualisation to help on predictive and prescriptive analytics is storytelling [96], which aims at presenting data with a narrative visualisation.

There are also visualisation tools specific for a given domain. For instance, in the field of climate-modelling, Lee *et al.* [113] developed a tool for visualisation of simulated Madden-Julian Oscillation, which is an important meteorological event that influences raining patterns from South America to Southeast Asia. The tool enables tracking of the event and its visualisation using Google Earth. In the area of computing networks management, Liao *et al.* [114] evaluated five approaches for visualisation of anomalies in large scale computer networks. Each method has its own applications depending on the specific type of anomaly to be visualised and the scale of the managed system. There are also solutions that provide means to visualise demographic information. Andrienko *et al.* [115] proposed interactive visual display for analysis of movement behaviour of people, vehicle, and animals. The visualisation tool displays the movement data, information about the time spent in a place, and the time interval from one place to another.

5.1. Open Challenges

There are many research challenges in the field of Big Data visualisation. First, more efficient data processing techniques are required in order to enable real-time visualisation. Choo and Park [106] appoint some techniques that can be employed with this objective, such as reduction of accuracy of results, coarsely processing of data points, compatible with the resolution of

the visualisation device, reduced convergence, and data scale confinement. Methods considering each of these techniques could be further researched and improved.

Cost-effective devices for large-scale visualisation is another hot topic for analytics visualisation, as they enable finer resolution than simple screens. Visualisation for management of computer networks and software analytics [116] are also areas that are attracting attention of researchers and practitioners for its extreme relevance to management of large-scale infrastructure (such as Clouds) and software, with implications in global software development, open source software development, and software quality improvements.

6. Business Models and Non-Technical Challenges

In addition to providing tools that customers can use to build their Big Data analytics solutions on the Cloud, models for delivering analytics capabilities as services on a Cloud have been discussed in previous work [3]. Sun *et al.* provide an overview of the current state of the art on the development of customised analytics solutions on customers' premises and elaborate on some of the challenges to enable analytics and analytics as a service on the Cloud. Some of the potential business models proposed in their work include:

- **Hosting customer analytics jobs in a shared platform:** suitable for an enterprise or organisation that has multiple analytics departments. Traditionally, these departments have to develop their own analytics solutions and maintain their own clusters. With a shared platform they can upload their solutions to execute on a shared infrastructure, therefore reducing operation and maintenance costs. As discussed beforehand, techniques have been proposed for resource allocation and scheduling of Big Data analytics tasks on the Cloud [67, 73].

- **A full stack designed to provide customers with end-to-end solutions:** appropriate for companies that do not have expertise on analysis. In this model, analytical service providers publish domain-specific analytical stream templates as services. The provider is responsible for providing the software stack and managing the resources necessary to perform the analyses. Customers who subscribe to the services just need to upload their data, configure the templates, receive models, and perform the proper model scoring.

- **Expose analytics models as hosted services:** analytics capabilities are hosted on the Cloud and exposed to customers as services. This model is proposed to companies that do not have enough data to make good predictions. Providers upload their models, which are consumed by customers via scoring services provided by the Cloud.

To make Big Data analytics solutions more affordable, Sun *et al.* [117] also propose cost-effective approaches that enable multi-tenancy at several levels. They discuss the technical challenges on isolating analytical artefacts. Hsueh *et al.* [93] discuss issues related to pricing and Service Level Agreements (SLAs) on a platform for personalisation in a wellness management platform built atop a Cloud infrastructure. Krishna and Varma [25] envision two types of services for Cloud analytics: (i) Analytics as a Service (AaaS), where analytics is provided to clients on demand and they can pick the solutions required for their purposes; and (ii) Model as a Service (MaaS) where models are offered as building blocks for analytics solutions.

Bhattacharya *et al.* [53] introduced IVOCA, a solution for providing managed analytics services for CRM. IVOCA provides functionalities that help analysts better explore data analysis tools to reduce the time to insight and improve the repeatability of CRM analytics. Also in the CRM realm, KXEN [118] offers a range or products for performing analytics, some of which can run on the Cloud. Cloud Prediction is a predictive analytics solution for Salesforce.com. With its Predictive Lead Scoring, Predictive Offers, and Churn Prediction, customers can leverage the CRM, mobile, and social data available in the Cloud to score leads based on which ones can create sales opportunities; create offers that have a higher likelihood to be accepted based on a prediction of offers and promotions; and gain insights into which customers a company is at risk of losing.

Cloud-enabled Big Data analytics poses several challenges in respect to replicability of analyses. When not delivered by a Cloud, analytics solutions are customer-specific and models often have to be updated to consider new data. Cloud solutions for analytics need to balance generality and usefulness. Previous work also discusses the difficulty of replicating activities of text analytics [119]. An analytical pathway is proposed to link business objectives to an analytical flow, with the goal of establishing a methodology that illustrates and possibly supports repeatability of analytical processes when using complex analytics. King [18], while discussing some of the problems in buying predictive analytics, provides a best practice framework based on five

steps, namely training, assessment, strategy, implementation, and iteration.

Chen *et al.* [120] envision an analytics ecosystem where data services aggregate, integrate, and provide access to public and private data by enabling partnerships among data providers, integrators, aggregators, and clients; these services are termed as DaaS. Atop DaaS, a range of analytics functionalities that explore the data services are offered to customers to boost productivity and create value. This layer is viewed as AaaS. Similar to the previously described work, they discuss a set of possible business models that range from proprietary, where both data and models are kept private, to co-developing models where both data and analytics models are shared among the parties involved in the development of the analytics strategy or services.

7. Gap Analysis

In business models where high-level analytics services may be delivered by the Cloud, human expertise cannot be easily replaced by machine learning and Big Data analysis [12]. Management should adapt to Big Data scenarios, and deal with challenges such as how to assist human analysts in gaining insights and how to explore methods that can help managers in making quicker decisions.

Application profiling is often necessary to estimate the costs of running analytics on a Cloud platform. Users need to develop their applications to target Cloud platforms; an effort that should be carried out only after estimating the costs of transferring data to the Cloud, allocating virtual machines, and running the analysis. This cost estimation is not a trivial task to perform in current Cloud offerings. Although best practices for using some data processing services are available [121], there should be tools that assist customers to estimate the costs and risks of performing analytics on the Cloud.

Data ingestion by Cloud solutions is often a weak point, whereas debugging and validating developed solutions is a challenging and tedious process. As discussed earlier, the manner analytics is executed on Cloud platforms resembles the batch job scenario: users submit a job and wait until tasks are executed to then download the results. Once an analysis is complete, they download sample results that are enough to validate the analysis task and after that perform further analysis. Current Cloud environments lack this interactive process, and techniques should be developed to facilitate interactivity and to include analysts in the loop by providing means to reduce their

time to insight. Systems and techniques that iteratively refine answers to queries and give users more control of processing are desired [122].

Furthermore, market research shows that inadequate staffing and skills, lack of business support, and problems with analytics software are some of the barriers faced by corporations when performing analytics [27]. These issues can be exacerbated by the Cloud as the resources and analysts involved in certain analytics tasks may be offered by a Cloud provider and may move from one customer engagement to another. In addition, based on survey responses, currently most analytics updates and scores of methods occur daily to annually; which can become an issue for analytics on stream data. Russom [27] also highlights the importance of advanced data visualisation techniques and advanced analytics—such as analysis of unstructured, large data sets and streams—to organisations in the next few years.

Chen *et al.* [123] foresee the emergence of what they termed as Business Intelligence and Analytics (BI&A) 3.0, which will require underlying mobile analytics and location and context-aware techniques for collecting, processing, analysing, and visualising large scale mobile and sensor data. Many of these tools are still to be developed. Moreover, moving to BI&A 3.0 will demand efforts on integrating data from multiple sources to be processed by Cloud resources, and using the Cloud to assist decisions by mobile device users.

8. Summary and Conclusions

The amount of data currently generated by the various activities of the society has never been so big, and is being generated in an ever increasing speed. This Big Data trend is being seen by industries as a way of obtaining competitive advantage over their competitors: if one business is able to make sense of the information contained in the data reasonably quicker, it will be able to get more costumers, increase the revenue per customer, optimise its operation, and reduce its costs. Nevertheless, Big Data analytics is still a challenging and time demanding task that requires expensive software, large computational infrastructure and effort.

Cloud computing helps in alleviating the latter two problems by providing resources on-demand with costs proportional to the actual usage. Furthermore, it enables infrastructures to be scaled up and down rapidly, adapting the system to the actual demand.

Although Cloud infrastructure offers such elastic capacity to supply computational resources on demand, the area of Cloud-supported analytics is still in its early days. In this paper, we discussed the key stages of analytics workflows, and surveyed the state-of-the-art of each stage in the context of Cloud-supported analytics. Surveyed work was classified in three key groups: Data Management (which encompasses data variety, data storage, data integration solutions, and data processing and resource management), Model Building and Scoring, and Visualisation and User Interactions. For each of these areas, ongoing work was analysed and key open challenges were discussed. This survey concluded with an analysis of business models for Cloud-assisted data analytics and other non-technical challenges.

Recurrent themes among the observed future work include (i) the development of standards and APIs enabling users to easily switch among solutions and (ii) the ability of getting the most of the elasticity capacity of the Cloud infrastructure. The latter includes expressive languages that enable users to describe the problem in simple terms while decomposing such high-level description in highly concurrent subtasks and keeping good performance efficiency even for large numbers of computing resources. If this can be achieved, the only limitations for an arbitrary short processing time would be market issues, namely the relation between the cost for running the analytics and the financial return brought for the obtained knowledge.

References

[1] F. Schomm, F. Stahl, G. Vossen, Marketplaces for data: An initial survey, SIGMOD Record 42 (1) (2013) 15–26.

[2] P. S. Yu, On mining big data, in: J. Wang, H. Xiong, Y. Ishikawa, J. Xu, J. Zhou (Eds.), Web-AgeInformation Management, Vol. 7923 of Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, 2013, p. XIV.

[3] X. Sun, B. Gao, Y. Zhang, W. An, H. Cao, C. Guo, W. Sun, Towards delivering analytical solutions in Cloud: Business models and technical challenges, in: IEEE 8th International Conference on e-Business Engineering (ICEBE 2011), IEEE Computer Society, Washington, USA, 2011, pp. 347–351.

[4] 'Big Data' has big potential to improve Americans' lives, increase economic opportunities, Committee on Science, Space and Technology (April 2013).
URL http://science.house.gov/press-release

[5] Prime minister joins sir ka-shing li for launch of £90m initiative in big data and drug discovery at Oxford, http://www.cs.ox.ac.uk/news/639-full.html (May 2013).

[6] bigdata@csail, http://bigdata.csail.mit.edu/.

[7] The Intel science and technology center for big data, http://istc-bigdata.org.

[8] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computing Systems 25 (6) (2009) 599–616.

[9] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, M. Zaharia, Above the clouds: A Berkeley view of Cloud computing, Technical report UCB/EECS-2009-28, Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, USA (February 2009).

[10] Attention, shoppers: Store is tracking your cell, New York Times.
URL http://www.nytimes.com/2013/07/15/business/attention-shopper-stores-are-tracking-your-cell.html

[11] Unlocking game-changing wireless capabilities: Cisco and SITA help Copenhagen Airport develop new services for transforming the passenger experience, Customer case study, CISCO (2012).
URL http://www.cisco.com/en/US/prod/collateral/wireless/c36_696714_00_copenhagen_airport_cs.pdf

[12] A. McAfee, E. Brynjolfsson, Big data: The management revolution, Harvard Business Review (2012) 60–68.

[13] B. Franks, Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics, 1st Edition, Wiley and SAS Business Series, Wiley, 2012.

[14] G. Bell, T. Hey, A. Szalay, Beyond the data deluge, Science 323 (5919) (2009) 1297–1298.

[15] T. H. Davenport, J. G. Harris, R. Morison, Analytics at Work: Smarter Decisions, Better Results, Harvard Business Review Press, 2010.

[16] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM 39 (11) (1996) 27–34.

[17] I. H. Witten, E. Frank, M. A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd Edition, Morgan Kaufmann, 2011.

[18] E. A. King, How to buy data mining: A framework for avoiding costly project pitfalls in predictive analytics, DMReview 15 (10).

[19] T. H. Davenport, J. G. Harris, Competing on Analytics: The New Science of Winning, Harvard Business Review Press, 2007.

[20] R. L. Grossman, What is analytic infrastructure and why should you care?, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 5–9. doi:10.1145/1656274.1656277.

[21] H. Wang, Integrity verification of Cloud-hosted data analytics computations, in: 1st International Workshop on Cloud Intelligence (Cloud-I 2012), ACM, New York, NY, USA, 2012.

[22] D. J. Abadi, Data management in the cloud: Limitations and opportunities, IEEE Data Engineering Bulletin 32 (1) (2009) 3–12.

[23] S. Sakr, A. Liu, D. Batista, M. Alomari, A survey of large scale data management approaches in cloud environments, IEEE Communications Surveys Tutorials 13 (3) (2011) 311–336. doi:10.1109/SURV.2011.032211.00087.

[24] D. S. Katz, S. Jha, M. Parashar, O. Rana, J. B. Weissman, Survey and analysis of production distributed computing infrastructures, CoRR abs/1208.2649.

[25] P. R. Krishna, K. I. Varma, Cloud analytics: A path towards next generation affordable BI, White paper, Infosys (2012).

[26] D. Jensen, K. Konkel, A. Mohindra, F. Naccarati, E. Sam, Business analytics in the Cloud, White paper IBW03004-USEN-00, IBM (April 2012).

[27] P. Russom, Big data analytics, TDWI best practices report, The Data Warehousing Institute (TDWI) Research (2011).

[28] P. Zikopoulos, C. Eaton, P. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data, McGraw-Hill Companies, Inc., 2012.

[29] PivotLink AnalyticsCLOUD, http://www.pivotlink.com/products/analyticscloud.

[30] X. Zhang, E. Zhang, B. Song, F. Wei, Towards building an integrated information platform for eco-city, in: 7th International Conference on e-Business Engineering (ICEBE 2010), 2010, pp. 393–398.

[31] J. K. Laurila, D. Gatica-Perez, I. Aad, J. Blom, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, M. Miettinen, The mobile data challenge: Big data for mobile computing research (2012).
URL `http://research.nokia.com/files/public/MDC2012_Overview_LaurilaGaticaPerezEtAl.pdf`

[32] A. Iosup, A. Lascateu, N. Tapus, CAMEO: enabling social networks for massively multiplayer online games through continuous analytics and cloud computing, in: 9th Annual Workshop on Network and Systems Support for Games (NetGames 2010), 2010, pp. 1–6. `doi:10.1109/NETGAMES.2010.5679650`.

[33] C. Wang, K. Schwan, V. Talwar, G. Eisenhauer, L. Hu, M. Wolf, A flexible architecture integrating monitoring and analytics for managing large-scale data centers, in: 8th ACM International Conference on Autonomic Computing, ICAC '11, ACM, New York, USA, 2011, pp. 141–150.

[34] D. Fisher, R. DeLine, M. Czerwinski, S. Drucker, Interactions with big data analytics, Interactions 19 (3) (2012) 50–59. `doi:10.1145/2168931.2168943`.

[35] S. Ghemawat, H. Gobioff, S.-T. Leung, The google file system, in: 9th ACM Symposium on Operating Systems Principles (SOSP 2003), ACM, New York, USA, 2003, pp. 29–43.

[36] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. S. Sarma, R. Murthy, H. Liu, Data warehousing and analytics infrastructure at Facebook, in: Proceedings of the 2010 international conference on Management of data, SIGMOD '10, ACM, New York, NY, USA, 2010, pp. 1013–1020. doi:10.1145/1807167.1807278.

[37] R. Ananthanarayanan, K. Gupta, P. Pandey, H. Pucha, P. Sarkar, M. Shah, R. Tewari, Cloud analytics: Do we really need to reinvent the storage stack?, in: Conference on Hot Topics in Cloud Computing (HotCloud 2009), HotCloud'09, USENIX Association, Berkeley, USA, 2009.

[38] F. Schmuck, R. Haskin, GPFS: A shared-disk file system for large computing clusters, in: 1st Conference on File and Storage Technologies (FAST'02), Monterey, USA, 2002, pp. 231–244.

[39] W. Tantisiriroj, S. W. Son, S. Patil, S. J. Lang, G. Gibson, R. B. Ross, On the duality of data-intensive file system design: reconciling HDFS and PVFS, in: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis, SC '11, ACM, New York, NY, USA, 2011, pp. 67:1–67:12. doi:10.1145/2063384.2063474.

[40] P. H. Carns, W. B. L. III, R. B. Ross, R. Thakur, PVFS: A parallel file system for linux clusters, in: 4th Annual Linux Showcase and Conference (ALS 2000), ALS '00, USENIX, Atlanta, USA, 2000, pp. 317–328.

[41] J. Cohen, B. Dolan, M. Dunlap, J. M. Hellerstein, C. Welton, MAD skills: new analysis practices for big data, Proc. VLDB Endow. 2 (2) (2009) 1481–1492.

[42] J. Kobielus, In-database analytics: The heart of the predictive enterprise, Technical report, Forrester Researc, Inc., Cambridge, USA (Nov. 2009).

[43] Amazon redshift, http://aws.amazon.com/redshift/.

[44] Amazon data pipeline, http://aws.amazon.com/datapipeline/.

[45] Amazon Elastic MapReduce (EMR), http://aws.amazon.com/elasticmapreduce/.

[46] G. DeCandia, D. Hastorun, M. Jampani, G. Kakulapati, A. Lakshman, A. Pilchin, S. Sivasubramanian, P. Vosshall, W. Vogels, Dynamo: Amazon's highly available key-value store, SIGOPS Operating Systems Review 41 (6) (2007) 205–220.

[47] N. Leavitt, Will NoSQL databases live up to their promise?, Computer 43 (2) (2010) 12–14.

[48] J. Han, H. E, G. Le, J. Du, Survey on NoSQL database, in: 6th International Conference on Pervasive Computing and Applications (ICPCA 2011), IEEE, Port Elizabeth, South Africa, 2011, pp. 363–366.

[49] R. Hecht, S. Jablonski, NoSQL evaluation—a use case oriented survey, in: International Conference on Cloud and Service Computing (CSC 2011), IEEE, Hong Kong, 2011, pp. 336–341.

[50] Sap hana one, http://www.saphana.com/community/solutions/cloud-info (2013).

[51] Birst Inc., http://www.birst.com.

[52] P. Deepak, P. M. Deshpande, K. Murthy, Configurable and extensible multi-flows for providing analytics as a service on the cloud, in: 2012 Annual SRII Global Conference (SRII 2012), 2012, pp. 1–10. doi:10.1109/SRII.2012.11.

[53] I. Bhattacharya, S. Godbole, A. Gupta, A. Verma, J. Achtermann, K. English, Enabling analysts in managed services for CRM analytics, in: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2009), KDD 2009, ACM, New York, USA, 2009, pp. 1077–1086. doi:10.1145/1557019.1557136.

[54] D. Habich, W. Lehner, S. Richly, U. Aßmann, Using cloud technologies to optimize data-intensive service applications, in: IEEE CLOUD, 2010, pp. 19–26.

[55] Datadirect cloud, http://cloud.datadirect.com/ (2013).

[56] R. Feldman, Techniques and applications for sentiment analysis, Communications of the ACM 56 (4) (2013) 82–89.

[57] J. Dean, S. Ghemawat, MapReduce: Simplified data processing on large clusters, Communications of the ACM 51 (1).

[58] Apache Hadoop, http://hadoop.apache.org.

[59] R. S. Barga, J. Ekanayake, W. Lu, Project daytona: Data analytics as a cloud service, in: A. Kementsietsidis, M. A. V. Salles (Eds.), International Conference of Data Engineering (ICDE 2012), IEEE Computer Society, 2012, pp. 1317–1320.

[60] Infochimps cloud overview, http://www.infochimps.com/infochimps-cloud/overview/.

[61] Windows Azure HDInsight, http://www.windowsazure.com/en-us/documentation/services/hdinsight/.

[62] K.-H. Lee, Y.-J. Lee, H. Choi, Y. D. Chung, B. Moon, Parallel data processing with MapReduce: A survey, SIGMOD Record 40 (4) (2011) 11–20.

[63] H. Herodotou, S. Babu, Profiling, what-if analysis, and cost-based optimization of mapreduce programs, Proceedings of the VLDB Endowment 4 (11) (2011) 1111–1122.

[64] Z. Guo, G. Fox, Improving mapreduce performance in heterogeneous network environments and resource utilization, in: 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid 2012), IEEE, 2012, pp. 714–716.

[65] A. Cuzzocrea, I.-Y. Song, K. C. Davis, Analytics over large-scale multidimensional data: the big data revolution!, in: Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP, DOLAP '11, ACM, New York, NY, USA, 2011, pp. 101–104. doi:10.1145/2064676.2064695.

[66] H. Herodotou, H. Lim, G. Luo, N. Borisov, L. Dong, F. B. Cetin, S. Babu, Starfish: A self-tuning system for big data analytics, in: In CIDR, 2011, pp. 261–272.

[67] G. Lee, B.-G. Chun, R. H. Katz, Heterogeneity-aware resource allocation and scheduling in the Cloud, in: 3rd USENIX conference on Hot topics in Cloud computing (HotCloud'11), HotCloud'11, USENIX Association, Berkeley, USA, 2011.

[68] Y. Chen, S. Alspaugh, R. Katz, Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads, Proceedings of the VLDB Endowment 5 (12) (2012) 1802–1813.

[69] G. C. Fox, Large scale data analytics on clouds, in: Proceedings of the Fourth International Workshop on Cloud Data Management (CloudDB 2012), ACM, 2012, pp. 21–24.

[70] D. Borthakur, J. Gray, J. S. Sarma, K. Muthukkaruppan, N. Spiegelberg, H. Kuang, K. Ranganathan, D. Molkov, A. Menon, S. Rash, R. Schmidt, A. Aiyer, Apache hadoop goes realtime at Facebook, in: ACM SIGMOD International Conference on Management of Data (SIGMOD '11), ACM, New York, USA, 2011, pp. 1071–1080.

[71] Y. Chen, S. Alspaugh, D. Borthakur, R. Katz, Energy efficiency for large-scale MapReduce workloads with significant interactive analysis, in: 7th ACM European Conference on Computer Systems (EuroSys '12), ACM, New York, USA, 2012, pp. 43–56.

[72] A. Balmin, K. Beyer, V. Ercegovac, J. M. F. Ozcan, H. Pirahesh, E. Shekita, Y. Sismanis, S. Tata, Y. Tian, A platform for eXtreme Analytics, IBM Journal of Research and Development 57 (3–4) (2013) 4:1–4:11.

[73] M. Rahman, X. Li, H. Palit, Hybrid heuristic for scheduling data analytics workflow applications in hybrid Cloud environment, in: IEEE International Symposium on Parallel and Distributed Processing Workshops and Phd Forum (IPDPSW), 2011, pp. 966–974.

[74] X. Li, R. N. Calheiros, S. Lu, L. Wang, H. Palit, Q. Zheng, R. Buyya, Design and development of an adaptive workflow-enabled spatial-

temporal analytics framework, in: 2012 IEEE 18th International Conference on Parallel and Distributed Systems, IEEE Computer Society, Singapore, 2012, pp. 862–867.

[75] H. Kasim, T. Hung, E. F. T. Legara, K. K. Lee, X. Li, B.-S. Lee, V. Selvam, S. Lu, L. Wang, C. Monterola, V. Jayaraman, Scalable complex system modeling for sustainable city (May 2013).

[76] H. Kim, S. Chaudhari, M. Parashar, C. Marty, Online risk analytics on the cloud, in: 9th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGrid 2009), IEEE Computer Society, Washington, USA, 2009, pp. 484–489.

[77] Google App Engine, http://developers.google.com/appengine/.

[78] N. Chohan, A. Gupta, C. Bunch, K. Prakasam, Hybrid Cloud support for large scale analytics and web processing, in: 3rd USENIX Conference on Web Application Development (WebApps 2012), Boston, USA, 2012.

[79] C. Bunch, N. Chohan, C. Krintz, J. Chohan, J. Kupferman, P. Lakhina, Y. Li, Y. Nomura, An evaluation of distributed datastores using the appscale cloud platform, in: 3rd IEEE International Conference on Cloud Computing (Cloud 2010), IEEE Computer Society, Washington, USA, 2010, pp. 305–312.

[80] G. Jung, N. Gnanasambandam, T. Mukherjee, Synchronous parallel processing of big-data analytics services to optimize performance in federated clouds, in: IEEE 5th International Conference on Cloud Computing (Cloud 2012), 2012, pp. 811–818.

[81] J. Chang, K. T. Lim, J. Byrne, L. Ramirez, P. Ranganathan, Workload diversity and dynamics in big data analytics: implications to system designers, in: Proceedings of the 2nd Workshop on Architectures and Systems for Big Data, ASBD '12, ACM, New York, NY, USA, 2012, pp. 21–26. doi:10.1145/2379436.2379440.

[82] R. N. Calheiros, C. Vecchiola, D. Karunamoorthy, R. Buyya, The Aneka platform and qos-driven resource provisioning for elastic applications on hybrid clouds, Future Generation Computer Systems 28 (6) (2012) 861–870.

[83] Q. Chen, M. Hsu, H. Zeller, Experience in continuous analytics as a service (CaaaS), in: 14th International Conference on Extending Database Technology, EDBT/ICDT '11, ACM, New York, USA, 2011, pp. 509–514.

[84] G. Yunwen, W. Shaochun, Y. Bowen, L. Jiazheng, The methods of data prefetching based on user model in cloud computing, in: Proceedings of the 2011 International Conference on Internet of Things, ITHINGSCPSCOM '11, IEEE Computer Society, Washington, DC, USA, 2011, pp. 463–466. doi:10.1109/iThings/CPSCom.2011.82.

[85] S. R. Upadhyaya, Parallel approaches to machine learning–a comprehensive survey, Journal of Parallel Distributed Computing 73 (3) (2013) 284–292.

[86] Apache Mahout, http://mahout.apache.org.

[87] B. Huang, S. Babu, J. Yang, Cumulon: Optimizing statistical data analysis in the Cloud, in: 2013 ACM SIGMOD International Conference on Management of Data SIGMOD '13, SIGMOD '13, ACM, New York, USA, 2013, pp. 1–12.

[88] A. Guazzelli, K. Stathatos, M. Zeller, Efficient deployment of predictive analytics through open standards and Cloud computing, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 32–38.

[89] Zementis – adaptive decision technology, http://www.zementis.com (2012).

[90] Google Prediction API, https://developers.google.com/prediction/.

[91] A. Kumar, F. Niu, C. Ré, Hazy: Making it easier to build and maintain big-data analytics, Communications of the ACM 56 (3) (2013) 40–49.

[92] A. Guazzelli, M. Zeller, W.-C. Lin, G. Williams, PMML: An open standard for sharing models, The R Journal 1 (1) (2009) 60–65.

[93] P.-Y. S. Hsueh, R. J. Lin, M. J. Hsiao, L. Zeng, S. Ramakrishnan, H. Chang, Cloud-based platform for personalization in a wellness management ecosystem: Why, what, and how, in: 6th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2010), 2010, pp. 1–8.

[94] IBM SmartCloud Enterprise, http://www-935.ibm.com/services/us/en/cloud-enterprise/ (2012).

[95] J. Davey, F. Mansmann, J. Kohlhammer, D. Keim, The future internet, Springer-Verlag, Berlin, Heidelberg, 2012, Ch. Visual analytics: towards intelligent interactive internet and security solutions, pp. 93–104.

[96] R. Kosara, J. Mackinlay, Storytelling: The next step for visualization, Computer 46 (5) (2013) 44–50.

[97] I. T. Tabor Communications, The UberCloud HPC Experiment: Compendium of Case Studies, Tech. rep. (2013).

[98] D. Fisher, I. Popov, S. M. Drucker, M. Schraefel, Trust me, i'm partially right: Incremental visualization lets analysts explore large datasets faster, in: SIGCHI Conference on Human Factors in Computing Systems (CHI 2012), CHI '12, ACM, New York, USA, 2012, pp. 1673–1682.

[99] B. Calder, J. Wang, A. Ogus, N. Nilakantan, A. Skjolsvold, S. McKelvie, Y. Xu, S. Srivastav, J. Wu, H. Simitci, J. Haridas, C. Uddaraju, H. Khatri, A. Edwards, V. Bedekar, S. Mainali, R. Abbasi, A. Agarwal, M. H. Fahim, M. H. Ikram, D. Bhardwaj, S. Dayanand, A. Adusumilli, M. McNett, S. Sankaran, K. Manivannan, L. Rigas, Windows Azure storage: A highly available Cloud storage service with strong consistency, in: 23rd ACM Symposium on Operating Systems Principles (SOSP 2011), SOSP '11, ACM, New York, NY, USA, 2011, pp. 143–157.

[100] S. Melnik, A. Gubarev, J. J. Long, G. Romer, S. Shivakumar, M. Tolton, T. Vassilakis, Dremel: Interactive analysis of web-scale datasets, Proceedings of the VLDB Endowment 3 (1–2) (2010) 330–339.

[101] F. B. Viegas, M. Wattenberg, F. van Ham, J. Kriss, M. McKeon, Manyeyes: a site for visualization at internet scale, IEEE Transactions on Visualization and Computer Graphics 13 (6) (2007) 1121–1128. doi:10.1109/TVCG.2007.70577.

[102] S. Lu, R. M. Li, W. C. Tjhi, K. K. Lee, L. Wang, X. Li, D. Ma, A framework for cloud-based large-scale data analytics and visualization: Case

study on multiscale climate data, in: 2011 IEEE Third International Conference on Cloud Computing Technology and Science (CloudCom 2011), IEEE Computer Society, Washington, USA, 2011, pp. 618–622. doi:10.1109/CloudCom.2011.95.

[103] Google Analytics, http://www.google.com/analytics/.

[104] Gooddata, http://www.gooddata.com (2013).

[105] S. Murray, Interactive Data Visualization for the Web, O'Reilly Media, 2013.

[106] J. Choo, H. Park, Customizing computational methods for visual analytics with big data, IEEE Computer Graphics and Applications 33 (4) (2013) 22–28.

[107] K. Reda, A. Febretti, A. Knoll, J. Aurisano, J. Leigh, A. Johnson, M. E. Papka, M. Hereld, Visualizing large, heterogeneous data in hybrid-reality environments, IEEE Computer Graphics and Applications 33 (4) (2013) 38–48.

[108] XSEDE, http://www.xsede.org/.

[109] SAP Crystal Solutions, http://www.crystalreports.com/.

[110] Cloud9 Analytics, http://www.cloud9analytics.com.

[111] Right90, http://www.right90.com.

[112] Salesforce, http://www.salesforce.com.

[113] T.-Y. Lee, X. Tong, H.-W. Shen, P. C. Wong, S. Hagos, L. R. Leung, Feature tracking and visualization of the Madden-Julian Oscillation in climate simulation, IEEE Computer Graphics and Applications 33 (4) (2013) 29–37.

[114] Q. Liao, L. Shi, C. Wang, Visual analysis of large-scale network anomalies, IBM Journal of Research and Development 57 (3/4) (2013) 13:1–13:12.

[115] G. Andrienko, N. Andrienko, S. Wrobel, Visual analytics tools for analysis of movement data, SIGKDD Explor. Newsl. 9 (2) (2007) 38–46. doi:10.1145/1345448.1345455.

[116] T. Menzies, T. Zimmermann, Software analytics: So what?, IEEE Software 30 (4) (2013) 31–37.

[117] X. Sun, B. Gao, L. Fan, W. An, A cost-effective approach to delivering analytics as a service, in: 19th IEEE International Conference on Web Services (ICWS 2012), Honolulu, USA, 2012, pp. 512–519.

[118] KXEN, http://www.kxen.com.

[119] L. Proctor, C. A. Kieliszewski, A. Hochstein, S. Spangler, Analytical pathway methodology: Simplifying business intelligence consulting, in: Annual SRII Global Conference (SRII 2011), 2011, pp. 495–500.

[120] Y. Chen, J. Kreulen, M. Campbell, C. Abrams, Analytics ecosystem transformation: A force for business model innovation, in: 2011 Annual SRII Global Conference (SRII 2011), SRII 2011, IEEE Computer Society, Washington, USA, 2011, pp. 11–20.

[121] P. Deyhim, Best practices for Amazon EMR, White paper, Amazon (2013).
URL `http://media.amazonwebservices.com/AWS_Amazon_EMR_Best_Practices.pdf`

[122] J. M. Hellerstein, R. A. A. Chou, C. Hidber, C. Olston, V. Raman, T. Roth, P. J. Haas, Interactive data analysis: the control project, Computer 32 (8) (1999) 51–59. `doi:10.1109/2.781635`.

[123] H. Chen, R. H. L. Chiang, V. C. Storey, Business intelligence and analytics: From big data to big impact, MIS Quarterly 36 (4) (2012) 1165–1188.