Publication pour la présente candidature LAFPT

Titre	Green and Distributed Architecture for Managing Big Data of Biodiversity
Auteurs	Idrissa Sarr ;Hubert Naacke; Ndiouma Bame; Ibrahima Gueye; **Samba Ndiaye**
Référence	Computing in Research and Development in Africa
Editeur	Springer
Pages	21 - 39
Année	2014
URL	**https://link.springer.com/chapter/10.1007/978-3-319-08239-4_2**
DOI	https://doi.org/10.1007/978-3-319-08239-4_2
Index	https://www.scopus.com/authid/detail.uri?authorId=6701604512
ISBN	978-3-319-08238-7
Encadreur	Non
Extrait d'une thèse	Non

Green and Distributed Architecture for Managing Big Data of Biodiversity

Authors   Authors and affiliations

Idrissa Sarr ✉ , Hubert Naacke, Ndiouma Bame, Ibrahima Gueye, Samba Ndiaye

Chapter
First Online: 02 September 2014

Abstract

The biodiversity term refers to the totality of genes, species, and ecosystems of a region or the globe. Biodiversity's impact on the human health and the ecosystem is without a doubt very significative. Therefore, the conservation of the biodiversity is becoming an international political and scientific issue since it may have a drawback on climate and the human health or survival. For a sustainable development perspective, several ongoing studies are conducted to analyze, predict, and face biodiversity changes. Such studies require a huge volume of data collected, stored, shared, and exploited intensively by researchers through the world by using web technologies and information systems as GEOBON, LifeWacth, GBIF, MosquitoMap. These systems handle an important amount of computing and database resources that must be optimized for avoiding maintaining useless resources while reducing considerably the energy usage. Actually, the goal of such optimization that we propose in this chapter is to adapt (increase or decrease) the number of resources for dealing with data of biodiversity based on the current load (or number of requests) while ensuring good performances. The benefits of doing so are manifold. First, it fits perfectly with the objectives of green computing or green IT that suggest to define computing systems efficiently and effectively with minimal or no impact on the environment. Second, it is well suited for African developing countries that encounter frequently energy problems and that miss enough funds to maintain complex infrastructures.

Keywords

Cloud Computing   Data Placement   Biodiversity Data
Global Biodiversity Information Facility   Place Replica

Copyright information

About this chapter

Buy this book on publisher's site

Reprints and Permissions

Green and Distributed Architecture for Managing Big Data of Biodiversity

Idrissa Sarr, Hubert Naacke, Ndiouma Bame, Ibrahima Gueye, and Samba Ndiaye

Abstract The biodiversity term refers to the totality of genes, species, and ecosystems of a region or the globe. Biodiversity's impact on the human health and the ecosystem is without a doubt very significative. Therefore, the conservation of the biodiversity is becoming an international political and scientific issue since it may have a drawback on climate and the human health or survival. For a sustainable development perspective, several ongoing studies are conducted to analyze, predict, and face biodiversity changes. Such studies require a huge volume of data collected, stored, shared, and exploited intensively by researchers through the world by using web technologies and information systems as GEOBON, LifeWacth, GBIF, MosquitoMap. These systems handle an important amount of computing and database resources that must be optimized for avoiding maintaining useless resources while reducing considerably the energy usage. Actually, the goal of such optimization that we propose in this chapter is to adapt (increase or decrease) the number of resources for dealing with data of biodiversity based on the current load (or number of requests) while ensuring good performances. The benefits of doing so are manifold. First, it fits perfectly with the objectives of green computing or green IT that suggest to define computing systems efficiently and effectively with minimal or no impact on the environment. Second, it is well suited for African developing countries that encounter frequently energy problems and that miss enough funds to maintain complex infrastructures.

I. Sarr (✉) • N. Bame • I. Gueye • S. Ndiaye
Université Cheikh Anta Diop, Fann - BP 5005, Dakar, Senegal
e-mail: Idrissa.Sarr@ucad.edu.sn; Ndiouma.Bame@ucad.edu.sn; Ibrahima.Gueye@ucad.edu.sn; Samba.Ndiaye@ucad.edu.sn

H. Naacke
Sorbonne Universités, UPMC Univ Paris 06, LIP6 Paris, France
e-mail: Hubert.Naacke@lip6.fr

1   Introduction

The biodiversity term refers to genetic variation, species variation, or ecosystem variation within a geographical area. It may have a great impact on the human health and the ecosystem. Therefore, the conservation of the biodiversity becomes an international and political issue in the last two decades. The main perspective behind this increasing attention is due to the fact that the biodiversity conservation is linked to sustainable development around the world. In this respect, a convention was entered on December 1993 between several countries to face challenges of the biodiversity conservation [18]. These challenges are manifold and are related to a various kind of aspects such as economic, health, science, and politic. A key issue when facing such challenges is to build and share a comprehensive inventory of all species of biodiversity in the world. The motivation of doing so is to follow up species, to understand their life conditions, and to forecast the growth and/or the depletion of their number.

However, these challenges require infrastructures and funds for defining and supporting good policies in order to collect and promote the sharing of the biodiversity data. Data may come from many institutions around the world with different structures and must be shared through the web. In this respect, a set of web platforms are built as GEOBON [10], LifeWacth [13], Global Biodiversity Information Facility (GBIF) [9], MosquitoMap [15]. These platforms bring together the diverse, stand-alone observation instruments and systems that track genetic resources, species, and ecosystems.

Furthermore, such platforms must remain available at any time while ensuring acceptable response time. In this respect, the architecture should be distributed and scalable, which requires powerful infrastructures and several resources. However, the resources of most of the existing platforms are set in a static (or by anticipation) way even though the workload fluctuates over time. In fact, the workload varies in such a way that there are some periods of high overall activity and some other more quiet periods. Hence, a static allocation may lead to a wastage since computational resources may be underused during quiet periods while they keep consuming power. Consequently, the static allocation by anticipation strategy does not permit to cope perfectly with the biodiversity convention that encourages the reduction of useless energy consumption.

The main goal of this chapter is to propose an elastic solution that has the ability to acquire and release resources on-demand in response to workload whose requirements fluctuate over time. The designed approach relies on the data access patterns and the characteristics of the GBIF web portal. Moreover, the proposed solution should be implemented on top of a cloud-based infrastructure that affords computing and storage capabilities with a low cost. The benefits of doing so are multiple. First, it fits perfectly with the objectives of green computing that suggests to define computing systems efficiently and effectively with minimal or no impact on the environment. Second, it is well suited for African developing countries that encounter frequently energy problems and that miss enough funds to maintain complex infrastructures.

Due to the large public we want to reach, we do not dive in deep into the technical aspects of the solution we describe. We refer readers to our work presented in [12] for details about the technical aspects. The rest of this paper is organized as follows. We portray in Sect. 2 the biodiversity and its conservation within African countries. We point out the challenges to face in such a context and we propose some solutions. In Sect. 3, we describe data of biodiversity of the GBIF and some of their use cases. Section 4 presents a green solution to deal with data biodiversity by defining first the data access pattern and highlighting the gain to be had when managing workload in a efficient way. We describe a solution for implementing a green management approach to deal with biodiversity workload in Sect. 5. A discussion of the overall advantages of the approach is given in Sect. 6 while Sect. 8 concludes.

2   Biodiversity in the Realm of African Countries

The conservation of biodiversity is considered as one of the big challenges and key issues of sustainable development since the Earth Summit held in Rio de Janeiro from 3 to 14 June 1992. Actually, the Convention on Biological Diversity as known as the Biodiversity Convention was entered on December 1993 with three main goals: (1) the conservation of the biodiversity; (2) a sustainable use of its components; and (3) a fair and equitable sharing of benefits arising from genetic resources [18]. Even though the ecosystems, species, and genes must be used for the benefit of humans, it is worth noting that natural resources are not infinite and require sustainable use. Hence, the usage of resources should be done without leading to a long-term decline or a dearth of biological diversity.

The specifications and requirements of the convention arouse many issues among which we point out: (1) the sharing of the results obtained from research and development related to genetic resources; (2) the coordination of a global directory of taxonomic expertise; (3) the education and public awareness; (4) the provision of financial resources; (5) the technical and scientific cooperation. These issues show all the complexity of conserving the biodiversity stability that requires several collaborations between scientifics, economists, politicians, sociologists, and so forth. In this respect, the issue of preserving biodiversity is attracting more and more interest, and mainly, in developing countries. Basically, such interests are translated by new policies and structures defined by African government to cope with the goals of the convention. For instance, it is quite impossible nowadays to see an African government without a ministry of Ecology or a ministry of sustainable development. Nevertheless, African countries still face problems to carry out with efficiency and success their policies due to many challenges that we highlight in the next subsection.

2.1   Challenges in African Developing Countries

African countries experiment recurrent problems related to biodiversity because of the fact that many genetic resources are overexploited. For instance, the shortage of good policies of regulation or appropriate tools to control the marine biodiversity leads to an overfishing and a depletion of fisheries that destroys marine mammals and entire ecosystem. As a consequence, illegal, unreported, and unregulated fishing is increasing in developing world since fishermen seek to avoid stricter rules in many places in response to shrinking catches and declining fish stocks. Moreover, in African countries, people partake to poaching for commercial gain, home consumption, and to face the lack of employment opportunities. The body parts of some animals are also in demand for traditional medicine and ceremonies. Furthermore, the dearth of water or rain in many sub-Saharan countries accelerates the disappearance of species and push people to move from an area to another. Such practices and human behaviors speed up the defaunation of forests, the reduction of animal populations, the emergence of zoonotic diseases, such as Ebola Virus, caused by transmission of highly variable retrovirus chains, and so on.

Therefore, it is obvious to observe that dealing with problems of biodiversity in Africa is challenging at many points. Hereafter, a short list of challenges we want to highlight in the context of this chapter.

- One of the main challenges is to perform a reliable and comprehensive inventory of all species biodiversity in Africa. The motivation of doing so is to follow-up species and to understand their environmental and life conditions. Such inventories can be used for predicting and modelling dynamic of species and will help to forecast the growth and/or the depletion of their number. Plus, it may monitor the utilization of genetic resources after they leave a country including by designating effective checkpoints at any stage of the value-chain. However, this inventory requires infrastructures and storage support for long term conservation and for a wide sharing. Such infrastructures are infrequent due to their expensiveness and if ever they exist, they are not enough efficient.
- Another challenge consist in defining strong, fair, and non-arbitrary rules procedures for regulating use of genetic resources or for protecting forest areas. Such procedures must be established for education and public awareness and for prior informed consent and mutually agreed terms. The problem of getting this goal is caused by the language barrier and the shortage of communication support for sensitizing every one of the drawback of overexploiting genetic resources. In fact, people of African countries do not speak the same language, and thus, each message should be translated in various dialects to reach people in remote rural areas. Moreover, some people resist for any change of their traditions or culture. Hence, such African realities make quite impossible or difficult to establish any consent and mutually agreed terms. Nevertheless, new technologies may help to communicate with such people through video messages and Internet. However, Internet and electricity are not present everywhere in Africa and remain entire issues.

- Last but not the least, African governments should create conditions to promote and encourage research contributing to biodiversity conservation and sustainable use. To this end, the creation of worldwide databases for sharing taxonomic expertise and semantic is a paramount issue. However, there is an increasing lack of policies or funds that do not encourage researchers to collect and share taxonomic. Even if some researchers keep working on biodiversity conversation by modelling the ecosystem and predicting its evolution, governments do not valorize their works by promoting them in a large public or rewarding them. Thus, researchers tend to go abroad where they may get infrastructures and supports to develop their ideas.

2.2   Solutions and Contributions

One may see that these challenges cited above cannot be faced without infrastructures and funds for defining and supporting good policies in order to collect and promote the sharing of the biodiversity data. Data may come from many institutions around the world with different structures and must be shared through the web. To this end, a set of web platforms or web portals are devised in developed countries such as GEOBON [10], LifeWacth [13], GBIF [9], MosquitoMap [15]. These platforms bring together the diverse, stand-alone observation instruments and systems that track genetic resources, species, and ecosystems. Moreover, these platforms can integrate biodiversity data with data on climate and other key parameters in order to fill gaps in taxonomic and biological information and speed up the pace at which information is collected and disseminated.

Since such platforms are designed for public and free use, thus, they can be used by African researchers for uploading or downloading data of biodiversity related to the continent and their research. In other words, these platforms can be an alternative solution of the challenges we cited in the previous section. Actually, African countries can rely to these information systems to face challenges caused by shortage funds required to build and maintain complex infrastructure for biodiversity data management.

Furthermore, such platforms must remain available at any time while ensuring acceptable response time. In this respect, the architecture should be distributed and scalable, which requires powerful infrastructures and several resources. However, the allocation of resources or their design is often done based on the prediction of maximum needs that the system could face in terms of computation (maximum peak load or simultaneous user requests) or in terms of storage (maximum volume of data). With the storage perspective issue, it is worth noting that resources are and remain generally appropriate and their extension are pretty simple. However, the allocation by anticipation of computational resources may lead to wastage since the appearance of the maximum load is infrequent. Therefore, allocated computational resources are often idle while they keep consuming power for cooling and requiring human resources for administrative tasks.

Consequently, the allocation by anticipation strategy does not permit to cope perfectly with the biodiversity convention that encourages the reduction of useless energy consumption. To deal with this issue, we propose an elastic solution that has the ability to acquire and release resources on-demand in response to workloads whose requirements fluctuate over time. The motivation of doing so is to optimize and/or reduce the number of resources in a dynamic fashion and to be able to get the objectives of "green computing" or "green IT."

Furthermore, African researchers need enough and sophisticated local infrastructures to store data they download or on which they want to work on. However, such infrastructures are not always available due to shortage of funds and lack of expertise. Hence, a solution to this problem is using a Cloud-based infrastructure to afford computing and storage capabilities with a low cost. In fact, Cloud computing allows companies to avoid upfront infrastructure costs, and focus on their businesses instead of infrastructure. Moreover, Cloud resources are usually shared by several users and are also dynamically adjusted per demand to meet fluctuating and unpredictable business needs. The Cloud has the advantage to maximize the use of computing powers thus reducing environmental damage since less power, air conditioning, rackspace, and so on is required for a set of functions. Plus, multiple users can access a single server to retrieve and update their data without purchasing licenses for different applications. This strategy fits well in the context of African countries where researchers have not enough financial support to bring into their team all skills or applications they need.

Briefly, the main contribution of this chapter is a combination of two mechanisms, namely, a cloud-based infrastructure and an elastic data management strategy for biodiversity. We apply our solution on the GBIF information system that we describe in next section before describing how we manage data of biodiversity in a green fashion.

3   Describing Data of Biodiversity: GBIF Case

The GBIF is an international organization created in 2001 to ease the collect, integration, and share of primary biodiversity data [9]. The GBIF database is hosted in Dannark. Data come from many institutions from around the world and are related to plants, animals, fungi, and microbes of the world. That is, African countries are invited to invest in such infrastructures by uploading data related to species they discover and identify. The GBIF mainly focuses on making scientific data on biodiversity available via the Internet using web services and can be used as a support of communication between politicians, policy makers, researchers, and the general public together. That is, the GBIF mission is to facilitate free and open access to biodiversity data worldwide for building and catalyzing sustainable development. Another goal of the GBIF is to promote participation and working through partners for collecting biodiversity data, building a computing and storage architecture to allow their integration, designing protocols, and standards to

insure scientific integrity and interoperability, in order to make analytical tools for improving decision-making. Actually, the GBIF objectives can be summarized as follow:

- an international project for inquiring the global primary biodiversity data (from genes to ecosystems);
- an information architecture that makes biodiversity data accessible and searchable through a single portal;
- a network of international organizations that play the role of information assistance and training for data providers and users;
- a reliable scientific data on which scientific analysis can be processed in order to establish trends for biodiversity management.

3.1 Data Type and Data Model

GBIF portal gathers several types of data, namely, primary data and metadata. Basically, primary data refer to species information (e.g., taxonomy and owner), the details of their observations (e.g., geographical position, country, and region) and documentation in the form of audio-visual (e.g., photographs, videos, and audio). However, metadata describes the primary data as well as specifying the details of their providers, their collections, and so forth.

Moreover, the database is made of collection that contains many occurrences of a species. An occurrence describes characteristics of a specimen and contains many taxonomic and geographical position fields. In addition, an occurrence represents the basic unit of information in the database. It is worth noting that a collection belongs to a provider and linked to a country. Furthermore, the database contains more than 400 million hits (occurrences) from more than 10,000 datasets and 560 providers. Data are structured by using a relational data model and stored in such a way that all the database is mapped into two kind of tables: a main relation that contains all information about occurrences, and a set of side tables (small sizes) containing essentially the metadata of occurrences. Finally, the number of data records available in the web portal increases day in and day out. For instance, from December 2008 to September 2013, the amount of data records has varied from 163 to 416 millions. Hence, this huge amount of resources on primary data of biodiversity data keeps attracting more users to the GBIF portal and leads to an increase of more complex workload.

3.2 Use Cases of GBIF Data

Use cases of biodiversity data are manifold, various, and involve a range of activities. A list of use cases of biodiversity data are available on the GBIF portal [9]. These use cases are set for analyzing the interaction between species,

the evolution of species, extinction risk, socioeconomic importance, the impacts of climate change, etc. In this chapter, we describe use cases of modelling the ecological niche, species co-occurrence, and data input.

Modelling Ecological Niche

Modelling the ecological niche of a species is a process of building a function or a model that uses environmental parameters in order to predict the probability of presence of a given species. This modelling aims at studying a species behavior such as its distribution, migration, threat of extinction with respect to environmental factors, and spatiotemporal dimensions. This modelling can be used in agriculture for assessing the impact of bees on a flower, or to figure out new orientations toward a novel kind of peanut seeds, etc. It is also interesting to model ecological niche in the realms of ecology and breeding in order to identify the areas to be protected for the survival of endangered species, and to measure the productivity of a given animal based on resources and coexistence of several other species.

As pointed out earlier, modelling the ecological niche requires a complex data set containing both data related to the studied species and climatic or environmental data of the geographical regions within which studies are conducted. Data related to species describe their characteristics while environmental data portray physical and chemical data such as temperature, precipitation, salinity, solar radiation, and their interactions with other species as the relationship predator/prey. The GBIF portal provides data related to species description and geographical presence while environmental parameters are obtained from other data sources such as BIOCLIM, WorldClim.

Species Co-occurrence

Calculations of co-occurrence are very common in data analysis of biodiversity. In order to model the interactions and/or dependencies between species co-occurrences are required and are used in community ecology. Co-occurrence modelling consists of proving that two or more species coexist in the same area during the same period. For this to be done, the density of two given species must be greater than a minimum value (threshold) within a time window and a space dimension. Basically, the co-occurrence is computed for each cell of a space. To calculate the co-occurrence of several species, the density of species are first determined within each cell before aggregating all densities for the entire space in order to figure out the co-occurrence of a couple of species.

Data Input

To ensure the quality of data available in a collection, updates are necessary. These updates are correction of existing data or insertion of new occurrences. The corrections (change or delete) are performed when errors are noted on the validity of

data records. These errors can impact the consistency of observations and results of analysis (e.g., aquatic species observed in a desert) or the taxonomic structure (e.g., an herbivore baobab). After each prospecting, new instances are collected and are inserted into the database. This involves treatments identification and validation to check the consistency of the field observations and the risk of redundancy with other data records.

Based on the above use cases, the data type, and the data model, it is obvious to note that data of biodiversity may be considered as Big data because of their volume, velocity, and variety. Dealing with Big data is seen as one of the major challenges of this decade in both computer science and information system. Therefore, the management of the biodiversity data stored and shared through the GBIF portal is a paramount issue. As pointed out in Sect. 2, data of biodiversity must be managed in a green fashion in order to fit into the biodiversity convention. We motivate in the next sections an approach for biodiversity data and we afterwards, present the architecture for that purpose. Finally, we describe solutions for facing workload related to biodiversity data.

4   Green Data Management of the Biodiversity Data

Aiming to manage huge biodiversity databases using a network of distributed machines, we face a data placement problem. We state the problem as follows: given a set of machines (each with its own data management capacity in terms of computation and storage), given a workload of data access requests that continuously arrive, we have to assess where (and when) to place the requested data such that all the requests operate in reasonable time, using the minimal number of machines. Knowing all the requests in advance would allow to find a near optimal data placement as well as a schedule to move data when necessary. However, this does not apply in our context where the workload is not entirely known a priori. Conversely, an arbitrary workload that does not reveal any access pattern would not benefit from any clever placement strategy: in such a case a random placement appears to be the best solution. In between these two types of workload (ranging from fully known to fully random) we investigate data placement strategies when the workload exhibits some regular patterns. Hopefully, biodiversity data access presents some specific patterns that we aim to exploit in order to manage data at lower cost (i.e., in a green fashion).

4.1   Data Access Pattern

The main pattern concerns the data popularity that relates to data taxonomy. The users generally target a predefined set of species because they have to decide about an action to preserve, understand, or forecast the evolution of such species.

In many cases, the users enter into a sequence of several requests about the same set of species for a period (from hours to days) that is long enough to deserve dynamic data management adjustment. We have also observed that some users involve several other users that in turn submit their own requests about the same species. We assume that such kind of sudden interest spread (or popularity peak) tends to occur at an increasing rate and an increasing intensity (i.e. higher peak level) as an effect of the increasing socialization level of the users. In other words, the users tend to crystallize around some data for some time, then switch their focus to another data that becomes the next collective target, and so on. The second pattern relies on the observation that the workload globally fluctuates over time. The high number of users does not actually smooth the aggregated workload level. There are still some periods of high overall activity and some other more quiet periods. This can be faced by either a static or dynamic resource provisioning at a global level in order to solve the green data management challenge.

4.2   Drawback of a Static-Based Resources Allocation

In fact, for most of the analysis done over the data biodiversity, a user has to download occurrences of species that are required for a given study. In addition, climatic and environmental data may also be required for a purpose of a study. In such a case, data have to be gathered or integrated through aggregation and join queries that are very expensive, mainly, for large partition of data. That is, the amount of data used in the management of biodiversity are tied to a heavy workload because of the number of potential users.

To handle both data storage and workload issues, the data are partitioned and distributed on different storage nodes. Rather, an important amount of data stored on one node may be related to a few number of species (plant or animal) that are most stressed or required based on their importance or role during a bounded time interval. That is, the workload of the data of the biodiversity fluctuates over time. Thus, even though data are distributed, the system may experiment bottlenecks at some resources. To avoid such bottlenecks appearance, a well-known solution is an over-provisioning approach that allocates a fixed high number of resources. This is done at the design step by allocating resources according to the highest expected load for each partition. In other words, knowing the size and the content of a partition, and how frequently it is accessed, one can estimate the number of resources to set for facing all the workload.

The problems with over-provisioning are that at a given time, the number of the fixed resources may be underutilized, which may lead to a waste of power. Another approach would be to set an average configuration, which consists of allocating resources according to the average load of the system. This approach seems to be better whatsoever, it can lead to a situation in which the allocated resources may not face a peak load higher than the average.

4.3 Benefits of an Elastic-Based Allocation

Using an elastic-based approach is a mechanism of dimensioning resources configuration regarding to the size of the workload. Since workload can be very huge, an elastic solution requires to have on-hands unbounded resources. Unbounded resources can be obtained from a cloud computing infrastructure that holds huge resources.

Cloud computing and/or elastic computing have emerged as successful paradigms for scaling up with a low cost. One of the main factors of this success is that elasticity aims at allowing resources variations in terms of amount while the system works. This strategy gives the possibility of having initially a minimal resources configuration that is increased or tuned, if need be, in order to ensure low latency or good scalability. In the opposite, resources may be reduced if the workload decreases and requires less than current fixed resources. In order to adapt such elasticity in the realm of a biodiversity data management, we need to design a couple of mechanisms that take into account data access pattern and characteristics of the biodiversity.

Many works have been done for studying elasticity with or without a cloud-based platform [4,5,7,11]. Among those solutions, we point out TransElas [11] that adapts the middleware size according to its load, and a solution described in [12] that permits to adapt the database layer size (processing capacities and storage) to the workload variations. The key idea of these works was to organize the database as partitions and to migrate partitions of a overloaded database to an underloaded one in order to ensure a bounded response time while minimizing the resources.

The migration process is conducted without interrupting the system or the current transactions processing. Several works have focused on this specific feature. In [6] the authors propose a pull-on-demand approach in which an index is updated when a migration decision is taken. The goal is to ensure pursuing workload processing either on the migration source or the destination one. Moreover, we can mention Slacker [1], a middleware solution that uses hot Back-Up tools to copy the database while allowing service continuing during this phase. The migration method is based on the available processing capacity in order to minimize the migration impacts response time.

Another dimension is the resources optimization of the data partitioning or replicas placement. Several works have been done to face this issue such as Schism [3] and Sword [16]. These approaches rely on current load level to place replicas. They use a graph partitioning algorithm to find a placement that prevents data distributed transactions. In the graph, vertices are tuples and edges the co-occurrence of tuples that appear in the same transaction. The main objectives of these works are to provide an improved throughput while providing fault tolerance and scalability for distributed OLTP data management systems. However, the power saving that is crucial in biodiversity context is missing.

More recently, the minimization of the migration cost is taken into account in the study [17]. This approach faces the operational costs minimization by adapting in

an elastic manner the system size depending on the tenant behavior. Moreover, the approach considers and quantifies performance degradation during a migration in both the source and destination servers.

In conclusion, elastic computation is gaining popularity over static provisioning computation and allows to regulate resources use in a flexible and extensible way. Therefore, it affords a great opportunity to avoid to unnecessarily tied-up resources for managing data. Moreover, it minimizes wasted storage space and power, and it exhibits low performance overhead, such that it does not lead to a significant longer latency that jeopardizes performance requirements or incurs extra costs. For this to be possible, we must implement elasticity in a highly dynamic and a transparent fashion, such that it hides all details and cope well with biodiversity requirements.

5   Elastic and Green Workload Management

Our goal is to propose an elastic approach to handle biodiversity data. Based on the data access pattern described in Sect. 4, the workload may fluctuate over short periods of time and reach a peak. Since peaks are infrequent, few resources are initially allocated for the data management. Later, when some data tend to be over-loaded, additional resources are added and the workload is distributed efficiently over all resources. In the opposite, when the workload decreases significantly, resources are released for keeping only the required amount of them. In terms of data management, adding a resource implies to assign a partition and its related workload to an ad hoc machine. Respectively, releasing a resource implies to retrieve a partition from the machine on which it was stored. Furthermore, in order to figure out data that have to be moved dynamically, we rely on data partitioning and replication mechanisms detailed below.

5.1   Biodiversity Data Partitioning and Replication

We exploit the partitioned nature of the data access pattern (described in Sect. 4), to partition the data. A user querying biodiversity data usually focuses on a small set of species. Users updating (or adding) data also target their access to some species. Therefore, we can logically partition the data according to the species.

More precisely, the species description of biodiversity data is well organized in a hierarchy called a taxonomy. We take the taxonomy into account to define the data partitioning such that partitions match the user queries as much as possible. The taxonomy has several levels. For instance, a family contains several genus, and a genus contains several species. A query accessing all the species of a given genus (or family) allows to define a genus-wide (resp. family-wide) partition. Once the partition is defined, we decide which resource to allocate to each partition, using a data placement strategy. Several partitions may reside at the same machine

as long as the corresponding workload does not exceed the machine capacity. The placement strategy differs depending on two types of access: (i) for write access, we must guaranty consistent data access when several users require common data simultaneously. Thus, we ensure that at any moment, partitions being updated concurrently are disjoint. (ii) For read only access, we may replicate a partition at several places when more resources are needed for accessing that partition fast enough.

Transactions are update operations (insertion, deletion, modification) processed on the GBIF data. We consider two kind of transactions, namely, single-partition and multi-partition transactions based on the number of partitions required by a transaction.

5.2 Handling Single-Partition Transactions

Obviously, a single-partition transaction is accessing only one partition. It consists of inserting either new species or modifying/deleting an existing one.

A straightforward method to process the transactions consistently is executing them in a serial way on a single physical machine. However, when the incoming transactions rate is increasing, a single machine may not be able to process all the transactions in due time. Therefore, we need to migrate some partitions in order to balance the load over several machines. To this end, we continuously track the load of each partition at every machine. We identify the data partitions on which the peaks occur or from which response time is lengthening. We identify the data partitions from which the peaks occur or from which response time is increasing. The migration algorithm operates in order to permit partitions located on the same machine to be used during the migration period.

The migration algorithm is decentralized to ensure scalability. That is, each machine facing a peak is responsible to migrate some of its partitions. A common protocol coordinates the allocation of the available machines that are candidate to accept new partitions.

Illustrative Example

Given a set of data partitions P_1, P_2, P_3, and P_4 stored on a storage resource DB_1 as depicted in Fig. 1. Each partition is represented by a square and we set the color of squares as follows: (1) light gray for an idle partition (no transaction accessing it), (2) black if the partition is heavy required (overloaded) and another color for an intermediate situation (neither idle nor overloaded). One can see that at 8 o'clock, all partitions of DB_1 are idle. However, at 9 o'clock, DB_1 is overloaded because of the partition P_4. In such a situation, the first step after identifying the partition causing the issue is to find where to move it. Consider DB_2 a second storage resource on which we have already three partitions P_5, P_6, and P_7 (see Fig. 2). Since DB_2 is
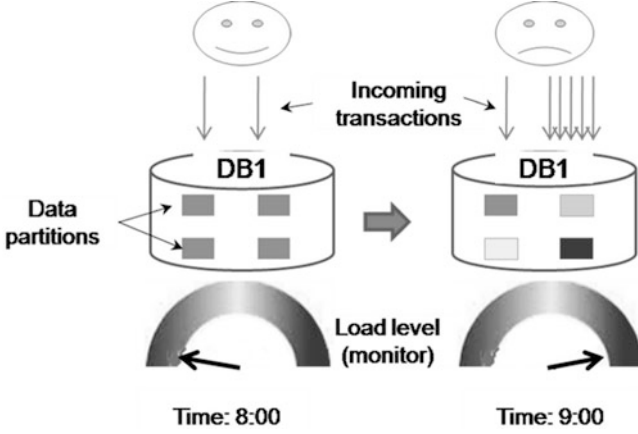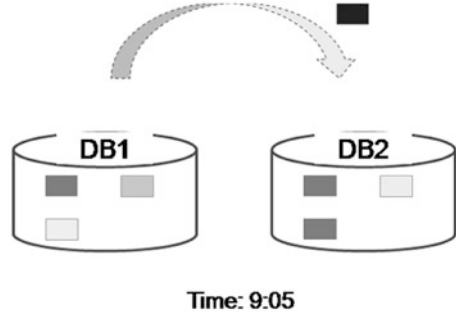
**Fig. 1** Load of DB$_1$ between 8:00 and 9:00

Fig. 2 Migrating P_4 from
DB$_1$ to DB$_2$

not overloaded (no black square), thus P_4 can be moved on it and all transactions
requiring P_4 are also moved on DB$_2$. It is worth noting that before migrating a
partition to another point, we assess if this latter will not be overloaded. That is why
in Fig. 3 everything is perfect after the migration completed. Finally, if ever there is
no available resource, a new DB instance will be created for receiving P_4 thanks to
the virtualization and elasticity capabilities of the cloud infrastructure we use.

5.3 Dealing with Multi-Partition Transactions

A multi-partition transaction accesses at least two partitions, and should be executed
on a single physical machine to avoid distributed transactions over a network, which
may break scalability performances [19]. Therefore, given a transaction requiring
many partitions, we first group the partitions on a single machine if ever they are not
already grouped. Precisely, groups are defined in such a way that each transaction is
performed on only one group. However, the grouping protocol may lead to gather a
high number of partitions into the same group. In such a case, migration, if any, will
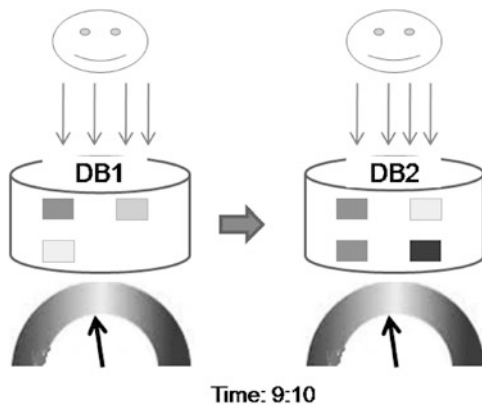
last more time while the related workload can reach rapidly a peak. One solution to avoid large groups is to limit the number of partitions that can be placed into a group.

Moreover, we track the load of each physical machine and migrate groups accordingly. However, when migrating groups of partitions instead of moving partitions one by one, the peaks may occur more frequently and avoid to get a good load balancing. With this in mind, it is important to be able to predict the occurrence of peak load and use this prediction to make the right choice when moving partitions or groups. Such prediction is beyond the scope of this chapter and we recommend reader to have a look on works described in [12], where we propose how using a social graph may help to forecast peak of workload.

6 Discussing About the Overall Gains of Our Solution

In this section, we summarize and discuss about the gains of our approach to manage workload of the biodiversity data. In short, we demonstrate through the paper that using elasticity and a cloud-bases infrastructure has the advantage to reduce cost and resources for managing data biodiversity with a green fashion. Rather, our solution has other advantages in terms of latency, scalability, parallelism, and data placement.

6.1 Controlling Parallelism

Queries for analyzing data of the biodiversity are complex and require heavy computation that may overpasses both the computational and storage capacity of a single machine. Hence, computing such queries must be done in a parallel manner on several machines. However, we do not want to parallelize as much as possible because this would create too much replicas, which require additional overhead for

data consistency control. Therefore, we design an algorithm to control the degree of parallelism for queries. To this end, we rely on a cost model for choosing an optimal plan to process a query. Basically, each query may be subdivided into many sub-queries that can be processed with several plans. The best plan is the one that reduces the amount of data to transfer from a machine to another one. Our algorithm coupled with the grouping strategy we described in the previous section ensure that parallelism is controlled. The motivation of doing so is that transferring data via a cloud infrastructure is costly. That is, reducing such transfers minimizes the financial cost.

6.2 Bounding the Response Time

The response time of a query must be bounded to satisfy the user requirement and depends on the current load status of the computational resources. Hence, to bound the response time, we need to track load of each resources when several users are simultaneously attempting to access them. Afterwards, the less loaded resource in terms of computation capabilities is chosen when a query can be processed by a set of resources. In addition, we record the capabilities of all resources in order to figure out the resource on which the response time is the lowest. The lowest time is estimated based on statistics, current loads, and capacities of available resources, and finally, the data transfer time.

6.3 Optimizing the Data Placement and Replication

Our grouping protocol described above coordinates the replication decisions as well as the data placement of the GBIF data. It prevents to overload resources that have enough and available capacities.

Moreover, replicas are created on the fly for an urgent need that corresponds to face a peak load. Once this load disappeared, the replica is deleted to reduce cost of maintaining consistency. Asynchronous replication is used to avoid lengthening response time. In most of the cases, users can be satisfied with slightly out-of-date replicas. Thus, synchronization of replicas is infrequent and planned for periods of low workload. The advantage of this infrequent synchronization is to reduce data transfers and allow saving money.

7 Related Work

The focus of this work is the management of large volume of data of the biodiversity. Many studies have been conducted to face issues of managing the biodiversity data [9, 14, 15]. Most of the proposed solutions are often oriented to a specific concept

such as a thematic-based approach [15] or a country-based one [2, 8]. For instance, thematic-oriented solutions permit the sharing of data by using thematic. Even though these solutions are frequently used and useful, data are highly tied to a small topics, and therefore, users who want information about other non-linked topics cannot be satisfied. Furthermore, some studies [14, 15] have focused on including analysis on biodiversity data. One of the well-known of such studies is MostiquoMap [18]. It is a very interesting approach that allows users to launch their analysis via a graphical user interface (GUI) from which results are displayed as a map after. However, this approach is tightly tied to MostiquoMap and may not be suited for various kind of biodiversity. Moreover, work described in [9,14,15] uses a centralized database for hosting all data and therefore, all requests are processed in a single point. Such approach has the drawback of jeopardizing the data availability as well as the scalability of the overall system. It is clear that centralization is not the best solution and particularly in a poor context where electricity is not always insured.

Our solution differs to the previous ones at many points. First, it distributes the data to where they are frequently used regardless of their origins and their themes. Users can also access the data regardless of their locations. Hence both scalability and availability are increased. Actually, our solution appears as a complement to previous solutions that do not directly share data sources or computing resources to meet all user needs. Second, our dynamic mechanism for data distribution and query processing ensures the scalability and integration of new features such as the ability to handle more complex queries and conducting analysis on data of biodiversity.

8 Conclusion and Perspectives

This chapter presents an approach to manage data of biodiversity, mainly, how the workload can be processed in a green fashion and to cope with the challenges aroused in the realm of African countries. In fact, data of the biodiversity may be considered as Big data because of their volume, velocity, and variety. Dealing with Big data is seen as one of the major challenges of this decade in both computer science and information system. Moreover, based on the convention of the biodiversity, several collaborations between scientifics, economists, politicians, sociologists, and so forth are required. Among the objectives of such collaborations, we cite the sharing of the results obtained from research and development related to genetic resources and the coordination of a global directory of taxonomic expertise. For this to be possible, expensive storage and computational infrastructures are required and they should be managed in such a way that the response time remains acceptable even though the workload is heavy. However, such infrastructures are not always available due to the dearth of funds and lack of expertise in most of the developing countries. To overcome these limits, we propose an elastic and cloud-based infrastructure solution that affords computing and storage capabilities with a low cost. In one hand, the cloud offers the advantage to optimize the use of

computing powers and multiple users can access a single server to retrieve and update their data without purchasing licenses for different applications. This strategy fits well in the context of African countries where researchers have not enough financial support to bring into their team all skills or applications they need. On the other hand, elastic computation allows to regulate resources use in a flexible and extensible way. Therefore, it affords a great opportunity to avoid to unnecessarily tied-up resources for managing data. Moreover, it minimizes wasted storage space and power, and it exhibits low performance overhead, such that it does not lead to a significant longer latency that jeopardizes performance requirements or incurs extra costs. For this to be possible, we must implement elasticity in a highly dynamic and a transparent fashion, such that it hides all details and cope well with biodiversity requirements. To this end, we propose some algorithms that track loads of resources and once one resource is overloaded, it processes a migration of partitions that cause the problem. The migration is done in such a way that distributed transactions are avoided and the load is well balanced.

Another issue that we do not take into account in this chapter is analytic workload that are made of complex queries and may last for a long while. Even though, analytic workload is crucial, we could not give more details due to a sake of presentation and space. We plan to portray how we face analytical workload in another paper.

References

1. S.K. Barker, Y. Chi, J.H. Moon et al., "Cut me some slack": latency-aware live migration for databases, in *International Conference on Extending Database Technology (EDBT)* (2012), pp. 432–443
2. Canadian BIF (2013), www.cbif.gc.ca
3. C. Curino, E.P.C. Jones, Z. Yang et al., Schism: a workload-driven approach to database replication and partitioning. VLDB Endow. **3**(1–2) 48–57 (2010)
4. C. Curino, E.P.C. Jones, S. Madden, Workload-aware database monitoring and consolidation, in *International Conference on Management of Data (SIGMOD)* (2011), pp. 313–324
5. S. Das, D. Agrawal, A. El Abbadi, ElasTraS: an elastic transactional data store in the cloud, in *International Conference on Hot topics in Cloud Computing* (2009)
6. A.J. Elmore, S. Das, D. Agrawal et al., Zephyr: live migration in shared nothing database for elastic cloud platforms, in *International Conference on Management of Data (SIGMOD)* (2011)
7. U.M. Farooq, R. Lui, A. Aboulnaga et al., Elastic scale-out for partition-based database systems, in *IEEE International Conference on Data Engineering (ICDE)* (2012), pp. 281–288
8. GBIF France (2013), www.gbif.fr
9. GBIF Secretary: GBIF data portal, GBIF web site (2013), data.gbif.org
10. GEOBON Web site (2013), www.earthobservations.org
11. I. Gueye, I. Sarr, H. Naacke, TransElas: elastic transaction monitoring for Web2.0 applications, in *Data Management in Cloud, Grid and P2P Systems* (2012), pp. 1–12
12. I. Gueye, I. Sarr, H. Naacke, Exploiting the social structure of online media to face transient heavy workload. In *The Sixth Intl. Conf. on Advances in Databases, Knowledge, and Data Applications*, IARIA (2014), pp. 51–58
13. LifeWatch Web Site (2013), www.lifewatch.com

14. Map of Life (2013), www.mappinglife.org
15. MosquitoMap (2014), www.mosquitomap.org
16. A. Quamar, K. Ashwin, A. Deshpande, SWORD: scalable workload-aware data placement for transactional workloads, in *International Conference on Extending Database Technology (EDBT)* (2013)
17. J. Schaffner, T. Januschowski, M. Kercher et al., RTP: robust tenant placement for elastic in-memory database clusters, in *International Conference on Management of Data (SIGMOD)* (2013), pp. 773–784
18. The Convention on Biological Diversity (2013), http://www.cbd.int/
19. A. Thomson, T. Diamond, S. Weng et al., Calvin: fast distributed transactions for partitioned database systems, in *SIGMOD* (2012)