Smart Data for ICT-based Water Management

Authors

Mariana MOCANU[1], Valentin CRISTEA[1], Ciprian DOBRE[1], Florin POP[1,*]

[1] Computer Science Department, University Politehnica of Bucharest, Romania

[1] Departamentul de Calculatoare, Universitatea Politehnica din București, Romania

Corresponding author

* Florin POP, florin.pop@cs.pub.ro

Abstract

Water is an essential, limited and sensitive life resource, and it is in focus of various persons or groups, from simple citizens to decision persons at country/world level, and, of course, also of scientists from different research fields. Water resource dynamic consequences exceed watersheds or water systems. Due to the support of new technologies, researches like people, water, and climate: adaptation and resilience in agricultural watersheds, developed a better understanding of the processes that link global-scale climate and socioeconomic drivers to regional-scale responses in land use decision-making, water quality, and water quantity. Recently, Cloud Computing emerged as the de facto state-of-the-art for data analytics. We require optimized platforms to co-locate data and computation and therefore mitigate the network bottleneck when moving data. However, as data may not be equally distributed across sites and since intermediate data are required to be aggregated to produce results, Cloud computing platforms may suffer severe performance degradation in such distributed settings. Thus, in our research activities we intend to address smart data extraction for water resource management, to explore new data distribution techniques and decision support systems that can co-operatively deal with distributed big data processing for single and multiple concurrent applications. Another challenging issue is to provide real-time analysis of shared and distributed data. While most real-time processing engines can efficiently benefit of the un-debatable performance of in-memory processing, they don't consider the data management during data processing (i.e. where to store the intermediate temporary data) or dependencies in-between processed data, which are common in environmental applications. In this case, mathematical models represent suitable instruments used in prediction and prognosis model for different parameters (i.e. water quality index), which are important for decision support systems for water resource management.

Procesarea inteligentă a datelor pentru managementul resurselor de apă

REZUMAT. Apa este o resursă esențială, limitată și sensibilă pentru viață, resursele de apă fiind în centrul atenției diferitelor persoane sau grupuri, de la simpli cetățeni la persoane de decizie la nivel de țară / nivel mondiali, un interes ridicat fiind arătat și de oameni de știință din diferite domenii de cercetare. Consecințele dinamice a managementului resurselor de apă au ca punct central depășirea capacității bazinelor de captare. Datorită suportului noilor tehnologii legate de adaptarea și capacitatea de adaptare în bazine hidrografice agricole, s-a dezvoltat o mai buna înțelegere a proceselor care se leagă de climă și a proceselor socio-economice în managementul resurselor de apă și oferă autorităților la scară globală răspunsuri ce se pot aplica la scară regională în luarea deciziilor cu privire la calitatea apei și cantitatea de apă folosită pentru consum. Recent, Cloud Computing a apărut ca un standard de facto pentru analiza de date. Necesitatea platformelor optimizate pentru a localiza date și a oferi resurse de calcul este o cerință impusă în serviciile de management global al resurselor de apă. Cu toate acestea, deoarece datele nu pot fi distribuite în mod egal și deoarece sunt necesare date intermediare să fie agregate pentru a produce rezultate corecte, platformele de calcul Cloud pot suferi o degradare de performanță severă. Astfel, în activitățile noastre de cercetare ne propunem să abordăm o extragere de date inteligentă pentru gestionarea resurselor de apă, pentru a explora noi tehnici de distribuție a datelor și a sistemelor de suport decizional, care pot coopera în prelucrarea datelor mari distribuite pentru aplicații concurente. O altă problemă dificilă este crearea unei analize în timp real a datelor partajate și distribuite. Cele mai multe platforme de procesare în timp real pot oferă performanțe atunci când datele sunt ținute în memorie, dar ele nu consideră managementul datelor în timpul procesării acestora sau a dependențelor în între datele prelucrate , care sunt comune în aplicații de mediu. In acest caz, modelele matematice reprezintă instrumente adecvate folosite în modelul de predicție și prognoză pentru diferiți parametri (indicele de calitate a apei), care sunt importante pentru sistemele de asistare a deciziilor în procesele de gestionare a resurselor de apă.

Keywords
water resources, smart data, big data, Cloud computing, decision support systems

Cuvinte-cheie
resurse de apă, procesare inteligentă a datelor, date masive, sisteme distribuite, sisteme de luare a deciziilor

Introduction

Small sensors and actuators are more and more used nowadays to extract knowledge about water-related problems. With the dawn of the Internet of Things (IoT), devices ranging from sensors monitoring the water pressure or leaks, to actuators, to even building, connect over the Internet. Infrastructures are being built to connect and collect data from the most diverse kind of devices monitoring water-related resources. Platforms such as InfluxData are constructed for information analytics, with a specialization on water management. Examples of IoT water-management applications include:

- Smart irrigation with IoT: Smart irrigation replaces existing irrigation controllers (which are just simple timers), with cloud enabled smart irrigation controllers that apply water based on plant need (i.e., type of crop) and weather. Moreover, with flow sensors and real-time alerts, property managers and landscape contractors can be alerted the second something goes awry, which if your site has any significant landscape at all, you know this can happen quite frequently. Examples of such systems: HydroPoint's WeatherTRAK® smart irrigation system (Khelifa et al., 2015).
- Smart water meters with IoT: A smart water meter (device) can collect usage data and communicate it wirelessly to the water utility company, where analytics software reports the results on a web site to view. Examples of such systems: One of the largest pilot programs of smart meters and related water management software platforms (a smart water management network) is in San Francisco. Water consumption is measured hourly and data is transmitted on a wireless basis to the utility four times a day. Both the utility and customers can track use. A pilot program in the East Bay Municipal Water District, which targets mostly single-family homes, provides a daily update of hour-by-hour consumption via a website. Consumers can be alerted, for example, by email or phone call, when water use exceeds a specified limit or when a meter indicates continuous running water for 24 hours. A customer can further view the data as it comes in, as well as compare their numbers with past use and city averages. The usage data should eventually result in alerts for leaks (by comparing how the readings in consecutive water meters) (Friess, 2013).
- Determining water demand in a city: One of the crucial challenges of water management as well as conservation in a city is to determine the amount of water that any city is going to utilize during the next day. This can be calculated to precision with the use of predictive analytics. Recently, IoT was employed for this purpose, where dedicated platforms keep a track on the history of water consumption in the city on any given day. Based on the historical data collected and analyzed by predictive analytics and combined with the consideration of special events, holidays, as well as the weather in that city, we can determine the amount of water that the entire population is going to consume in one day. The Internet of Things technology also helps in scheduling the maintenance as well as shutdown of pumps on a regular basis. There are optimization techniques which can beforehand convey to the residents of a city regarding the unavailability of water during any point of time. This helps the water regulation authorities in not only meeting the adequate water demands in a city; rather it also aids in the conservation of resources and energy.

In this paper, we analyze some of the decision factors when you are faced with decisions related to how to construct a water-management ICT support tool. The solutions presented in the first part of the paper are a collection of exiting models and technologies. In the second past, a Cloud-based application is presented. This application compute the Universal Water Quality Index (UWQI) (Boyacioglu, 2007).

Data integration, aggregation, and representation

The first decision relates to making decisions on the data formats and support to use, for the data you intend to collect. For water management, models can be derived from analysis and observation of the natural world (just by look at the water-related phenomenon). However, such models are prone to potential misunderstanding if they do not adhere to standards. Thus, a better approach is to rely on an open and integrated planning processes such as Integrated Water Resource Management (IWRM) (Voinov et al., 2008).

In water management, researchers and practitioners tend to agree that each case use best a tool or different model - it is simply up to the planner to select the best approach. In this sense, the Global Water Partnership, one of the largest forum crated around the IWRM concept, crated a set of policies and approaches they recommend to practitioners interested in the implementation of IWRM. Their recommendation includes references to a set of

Management Instruments, which are the proposed techniques to control water supply and demand. For these techniques, many models have been designed to facilitate integration between various aspects of catchment hydrology, including surface water, groundwater, vegetation, ecology, and even agricultural economics. Examples include NELUP (O'Callaghan, 1995), MIKE SHE (Refsgaard et al., 1995), and TOPOG (Vertessy et al., 1994). Such types of model are excellent for water resource assessments and impact on the environment, but in most cases, they do not link directly to the wider social, cultural and economic aspects of water management. Which is why researchers have proposed decision support systems (DSSs), as complementary tools to models. A DSS is a means of collecting data from many sources to inform a decision. Information can include experimental or survey data, output from models or, where data is scarce, and expert knowledge.

DSS tools and models were proposed in various studies about water monitoring/management (De Zwart, 1995), and are usually specifically tailored for one problem, to sustain the case being presented in each work. For example, diffuse of pollution from nutrients, namely nitrogen and phosphorus was presented in a vast study in (Munafo et al., 2005). As the article specifies, the number of chemicals released into surface water bodies is extremely large; their dynamics are complex and it is difficult to measure the global impact. The European inventory of existing chemical substance (EINECS) identified more than 100,000 chemicals, but there is not satisfactory knowledge of their routes of entry into surface waters yet. Furthermore, EINECS is likely to have underestimated the number of pollutants, for it does not consider all by-products deriving from physical, chemical, and biological degradation (Geiss et al., 1992). The management of non-point pollution of rivers and its prevention are priority factors in water monitoring and restoration programs.

The scientific community proposed many models for depicting the dynamics of pollutants coming from diffuse sources. In fact, most of them can be grouped into two broad categories: statistical models and physically based models. A major drawback of statistical or physically based models for non-point pollution is the large amount of data required both as input and for calibration and validation of the model. Other possible problems are long computing time, complexity related to the development of appropriate models, and the highly skilled operators required for using them. More recently, the potential non-point pollution index (PNPI) was proposed as s a GIS-based, watershed-scale tool designed using multi-criteria technique to pollutant dynamics and water quality (Munafo et al., 2005). The method for calculating PNPI follows an approach quite like the environmental impact assessment. The pressure exerted on water bodies by diffuse pollution coming from land units is expressed as a function of three indicators: land use, runoff, and distance from the river network. They are calculated from land use data, geological maps and a digital elevation model (DEM). The weights given to different land uses and to the three indicators were set per experts' evaluations and allow calculation of the value of the PNPI for each node of a grid representing the watershed; the higher the PNPI of the cell, the greater the potential impact on the river network.

Among the tools to support hydrological modelling and decision-making, Geographical Information System (GIS) is highly regarded as an important instrument for data management. For example, even when surface water and groundwater are modelled separately, GIS can support an integration between them (Facchi et al., 2004). For example, modelling software like Mike BASIN is selected often by different authors to model surface water. Groundwater models are also available are available in the ASM (Aquifer Simulation Model) software. When both surface water and groundwater need to be modeled together, both for quantity and quality evaluations, such tools (the complexity comes from the integration of the models these two provide) can be by means of a GIS, to support efficient data management. Such an approach was demonstrated in (Jain et al., 2004), where authors developed a process oriented distributed rainfall runoff model which used a GIS to generate model inputs in terms of land use, slope, soil and rainfall. This allowed the model to handle catchment heterogeneity.

Similarly, the GIS software ArcView, developed by ESRI, combines several capabilities for mapping systems along with the ability to analyze geographic locations and the information linked to those locations. A powerful feature of ArcView GIS is the ability to carry out mathematical and logical operations on spatial data. Furthermore, tabular data from Arcview dBASE files can be created or manipulated using Microsoft Excel, which is useful in facilitating the integration of ArcView with other software.

MIKE BASIN, developed by DHI software, is an extension of ArcView, which uses GIS information as a basis of a water resources evaluation (Hughes and Liu, 2008). Crucially, MIKE BASIN adds to ArcView the capability to deal with temporal data, in addition to the spatial data stored in the GIS. MIKE BASIN is a water resources management tool which is based on the basin-wide representation of water availability. Rivers and their main tributaries are represented mathematically by a network of branches and nodes. Nodes are point locations, where it is assumed that water enters or leaves the network through extractions, return flow and runoff. These may be confluences, diversions, locations where certain water activities occur (such as water offtake points for irrigation or

a water supply), or important locations where model results are required. Rainfall-runoff modelling can be carried out in MIKE BASIN using NAM (Nedbor Afstromnings Model), a lumped, conceptual rainfall-runoff model suitable for modelling rain-fall-run-off processes on the catchment scale. This can be used to simulate overland water flows, for example.

Aquifer Simulation Model for Microsoft Windows, is a complete two-dimensional groundwater flow and transport model. ASM include the instruments to model either confined and unconfined aquifers. For modelling an aquifer as a confined aquifer, the governing equations are based on transmissivity parameters, which are fixed because the saturated depth is fixed (when the water level in the aquifer drops below the confining layer, the saturated depth of the aquifer decreases, as does the transmissivity; thus, strictly speaking, the model is fundamentally flawed in this manner). For a steady-state model, the groundwater levels do not change once the solution has converged. Therefore, in such a model the transmissivity is effectively fixed, meaning the basic assumptions are still valid, however the data used to define the model should be based on measured or calibrated transmissivity and not on measured hydraulic conductivity. This also means that only steady-state analysis can be carried out with this model.

But the power of such modeling tools can be use when combined. As a pioneer case study, authors in (Ireson, 2006) proposed a methodology for loosely-coupling the MIKE BASIN with the ASM provided water models, and demonstrate a series of what-if scenarios for the effect of dams on the groundwater.

Collecting the data

In Europe, participation in water resource planning gained a new institutional stature with the Water Framework Directive (WFD). This calls for the active involvement of all interested parties in the implementation process and particularly in the production, revision, and updating of River Basin Management Plans (Article 14; Council of the European Communities, see (EC, 2000). Planning methods that combine public participation with decision-making functions are therefore increasingly in demand (EC, 2002). For example, several hydrography databases exist for the EU water studies that include rivers and lakes coverages. The catchments have been derived from a hierarchical river network, together with climate data provided for over 5k stations in all EU member states, collected by the monitoring agriculture with remote sensing (MARS) project (Vossen, 1995). The two main climatic variables are precipitation (average, maximum 24 h rainfall, number of rain days, average snowfall, number of snowfall and snow cover days) and temperature (average, maximum, minimum, absolute monthly maximum and minimum, number of frost days). Other climate attributes include, relative humidity, air pressure, atmospheric pressure, bright sunshine, evapotranspiration, wind speed and cloud cover.

Many more such initiatives were developed in the last years. The Waterkeeper Alliance, for example, developed programs (e.g., Riverkeeper, Lakekeeper, Baykeeper, and Coastkeeper) for ecosystem and water quality protection and enhancement, with major pilots in USA, Australia, India, Canada and the Russian Federation (Mohn, 2006). The URI Watershed Watch Program produces quality data from over 200 monitoring sites statewide (and citizens are encouraged to participate as active data readers). Produced and processed in certified laboratories, this information is used by the Rhode Island Department of Environmental Management for assessing the State's waters, as well as by municipal governments, associations, consulting firms and residents for more effective management of local resources. Similarly, Florida's LAKEWATCH program is one of the largest US lake monitoring programs in the nation with over 1800 trained citizens monitoring 600+ lakes, rivers and coastal sites in more than 40 counties. Volunteers take samples to collection sites located in 38 counties (Canfield et al., 2002).

Normally the use of water for productive activities is prohibited in the domestic distribution systems in many parts of the globe, but because these activities sustain in some places the rural poor, users withdraw water for unauthorized productive uses or alternatively water designated for irrigation is used to meet their domestic needs (Van der Hoek, 1999), leading to low availability and low quality of water. The use of "potable" water for all activities has become common, and other sources such as rainwater harvesting or grey-water re-use have been largely ignored in much of Latin America, for example (Restrepo, 2005). One factor that impedes decision making to improve water services in rural areas is the lack and inconsistency of information on water consumption, availability and quality (Roa et al., 2008). Without data, users cannot demonstrate causes of contamination and/or over exploitation of the resource, limiting their ability to lobby local authorities for improvements. Knowing water needs, water availability and the way human activities are affecting the resource, permits a diagnostic of overall watershed conditions, and the determination of priority sites for intervention.

In Romania authors in (Teodosiu et al., 2013) present a case study of how public participation, within the context of Integrated Water Resources Management (IWRM), promoted by promoted by the Global Water Partnership (GWP). IWRM is defined as "The process that promotes the coordinated development and

management of water, land and related sources to maximize the resultant economic and social welfare in an equitable manner, without compromising the sustainability of vital ecosystems" (GWP, 2000). The implementation of IWRM requires a participatory approach (Odendaal, 2002). It means that water management authorities should involve relevant stakeholders, such as representatives of water companies, industry, municipalities, agriculture, services, environmental protection agencies, non-governmental organizations (NGOs), universities and research institutions in planning, decision-making and implementation, instead of adopting a top-down approach (Casteletti et al., 2007). The importance of public participation (PP) in water management is also recognized by the European Commission through its Water Framework Directive (WFD, 2000/60/EC), which was the first directive that explicitly asks member states to inform and consult the public. Other directives, for example, on environmental assessments (2001/42/EC) and floods (FD, 2007/60/EC), have introduced similar requirements.

The implementation of these requirements is particularly challenging for new member states of the European Union (EU), many of them being post-communist countries. These countries are characterized by major environmental problems, and although the European requirements have been transposed into national legislation, practical application of PP is still lagging (Kremlis and Dusik, 2005). The governments of these new EU members rather give priority to the establishment of competitive markets and liberalization, while neglecting the development and empowerment of strong civil society representatives that would play active roles in the implementation of IWRM.

In Romania, besides the huge challenge of complying with the water quality standards of the WFD, there are serious issues to be addressed within the development of effective public participation. The case studies in (Teodosiu et al., 2013) show that the role of PP in dealing with these challenges is still limited. The first case shows that the traditional stakeholders, especially the water management authorities, still see PP as a simple formal requirement for the implementation of the WFD. Other stakeholders, especially NGOs and water users, feel the need for better representation and involvement, not only in public information and consultation activities, but also in the decision-making processes. In practice, as the case of formal participation in the development of river basin management plans shows, stakeholders are often very passive in reacting on plans. And, when stakeholders are engaged in an early stage of the planning process, as is shown in the case of active stakeholder involvement, authorities are reluctant to use the results.

For data collection, more recently people turned their attention towards what is called Participatory Sensing (Campbell et al., 2006). Unlike the traditional questionnaire-based collection processes, participatory sensing relies on electronic means widely available for collecting the data with the help of ordinary people. As mobile phones, have evolved from devices that are just used for voice and text communication, to advanced platforms that can capture and transmit a range of data types (image, audio, and location), the adoption of these increasingly capable devices by society has enabled a potentially pervasive sensing paradigm - participatory sensing. A coordinated participatory sensing system engages individuals carrying mobile phones to explore phenomena of interest using in situ data collection (Paulos et al., 2008). By enabling people to investigate previously difficult to observe processes with devices they use every day, participatory sensing brings the ideals of traditional community based data collection and citizen science to an online and mobile environment, while offering automation, scalability, and real-time processing and feedback (Cooper et al., 2007). In participatory sensing, individuals explicitly select the sensing modalities (they are in control of their privacy-related data) to use and what data to contribute to larger data collection efforts.

Processing large amount of data, its efficient and secure storage, data processing and sharing

The next step after deciding on the right models and tools to describe the problem at hand, is to consider how to process and extract useful knowledge out of large amounts of data potentially being captured and stored from water-related sensors. Several choices for runtime environment to help distribute the data analytics processing are presented below (our original analysis on the topic was previously published in (Dobre and Xhafa, 2014)). The hardware support of parallelism / concurrency varies from shared memory multicore, closely coupled clusters, and higher-latency (possibly lower bandwidth) distributed systems. The coordination (communication/synchronization) of the different execution units vary from threads (with shared memory on cores), MPI (message passing interface, between cores or nodes of a cluster), workflow or mash-ups linking services together, and the new generation of data intensive programming systems typified by Hadoop (implementing MapReduce) or Dryad. Short running threads can be spawned up in the context of persistent data in memory and have modest overhead (Fox et al., 2010). Short running processes (i.e., implemented as stateless services) are seen in Dryad and Hadoop. Also, various runtime platforms implement different patterns of operation. In Iteration-based platforms, the results of one stage are iterated many times. This is typical of most MPI

style algorithms. In Pipelining-based platforms, the results of one stage (e.g., Map or Reduce operations) are forwarded to another. This is functional parallelism typical of workflow applications. A (non-comprehensive) presentation of technologies in use today for Big Data processing is presented in Figure 1.

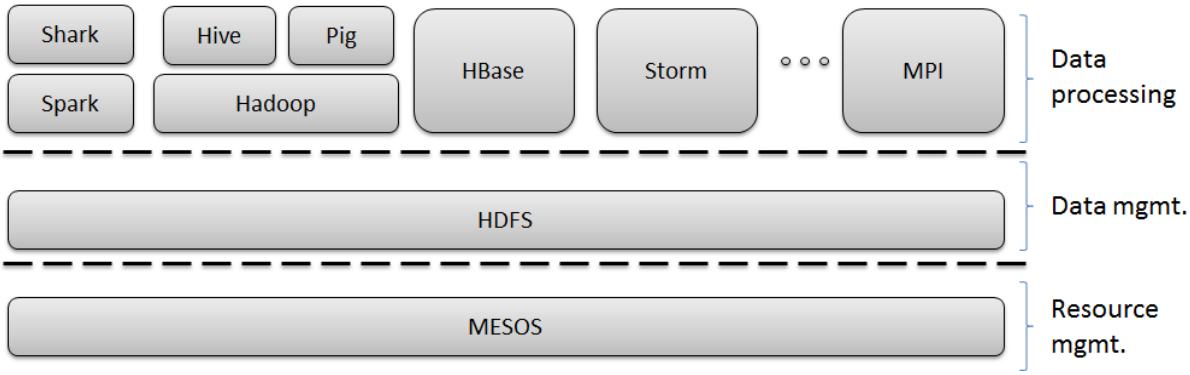

Figure 1: Example of an ecosystem of Big Data analysis tools and frameworks (Dobre & Xhafa, 2014)

In the mid-2000s we witnessed the first problems dealing with large volumes of data, like analyzing internet-scale of data or interpreting genomics data (the first "popular" HPC problems). Suddenly the High-Performance Computing community had problems to solve, where scalability, accuracy, large-scale data storage, and distributed matrix arithmetic became mainstream. This was an Era when people tackling such problems started creating parallel computing stacks, and MPI inarguably supported the initial growth of cluster computing. Even for data analytics related to assessing water management processes MPI proved its valuable support, due to its elegant support for general reductions (Camp et al., 2011). However, soon enough the scientific community wanted more, as MPI failed to deliver support for fault tolerance, and/or it failed to show the flexibility that later alternative tools (such as Hadoop or Dryad) brought. It took 14 years to go from MPI-2 to MPI-3, and even still it has a hardcoded in 32-bit limit throughout almost its entire API, limiting how many objects it can deal with at once without going through pointless but straightforward hoops. No wonder that the HPC community moved on.

Later, MapReduce (MR) emerged as an important programming model for large-scale data-parallel applications (Dean and Sanjay, 2008). MapReduce breaks a computation into small tasks that run in parallel on multiple machines, and scales easily to very large clusters of inexpensive commodity computers. The most popular open-source implementation of MapReduce is today Hadoop (Zaharia et al., 2008), and includes several specific components, such as its own file system, or support for fault tolerance and for scheduling in heterogeneous clusters. Due to its simplicity in design, no wonder that even today many projects relying on the use of computer tools for water-related data analytics rely on Hadoop in support for processing large volumes of sensed data (Zhang et al., 2015; Jach et al., 2015).

The next generation of HPC tools includes platforms such as Pig or Dryad. The problem with the MapReduce model is that it cannot be applied straightforward to all problems. The HPC community soon discovered that, although adequate for indexing, for problems from the realm of machine learning and data predictions it was not that easy to use. Thus, Pig (Olston et al., 2008) and later Hive (Thusoo et al., 2010) was developed on top of the MapReduce model to hide some of the complexity from the programmer, offering a limited hybridization of declarative and imperative programs and generalize SQL's stored-procedure model. Twister is another MapReduce extension, designed to support iterative MapReduce computations efficiently (Ekanayake et al., 2008) based on a publish/subscribe messaging infrastructure for communication and data transfers. Dryad is a general-purpose distributed execution engine for coarse-grain data-parallel applications (Isard et al., 2007), that allows fine control over the communication graph as well as the subroutines that live at its vertices. From these examples, Dryad is designed to scale from powerful multi-core single computers, through small clusters of computers, to data centers with thousands of computers. The Dryad execution engine handles all the difficult problems of creating a large distributed, concurrent application: scheduling the use of computers and their CPUs, recovering from communication or computer failures, and transporting data between vertices.

Finally, we are now in the moment when even such tools, designed to optimize the way data is handled and processed over novel database models (i.e., such as NoSQL and NewSQL), is simply not enough anymore. Data scientists want even more scalability and faster delivery of results from their tools, and so the early 2010s witnessed the development of the current wave of HPC tools: In-Memory Processing (or, sometimes called In-Memory Computing). Spark is among the pioneering framework that supports this processing model (Zaharia,

2010). In-Memory Computing may be defined as a solution that stores data in RAM, across a distributed system (cluster, cloud), and processes it in parallel. Spark provides two main abstractions for parallel programming: resilient distributed datasets and parallel operations on these datasets (invoked by passing a function to apply on a dataset). Resilient distributed datasets (RDDs) are read-only collections of objects partitioned across a set of machines that can be rebuilt if a partition is lost. Users can explicitly cache an RDD in memory across machines and reuse it in multiple MapReduce-like parallel operations. RDDs achieve fault tolerance through a notion of lineage: if a partition of an RDD is lost, the RDD has enough information about how it was derived from other RDDs to be able to rebuild just that partition. As per experiments (Zaharia et al., 2010), by making use extensively of memory storage (using the RDD abstractions) of cluster nodes, most of the operations Spark can outperform Hadoop by a factor of ten in iterative machine learning jobs, and can be used to interactively query a large dataset with sub-second response time. Other in-memory tools include examples such as Apache Ignite or SAP's HANA (Mazumder et al., 2016). Apache Ignite is an In-Memory Data Fabric that combines different components like in-memory data grid, in-memory computing grid and in-memory streaming into the same unique solution. SAP's HANA is an in-memory database that provides large data analysis and aggregation. It uses very large amounts of main memory, multi-core CPUs on multiple nodes in a cluster, and SDD storage, to improve the performance. Around such tools, projects already appear that make use in-memory processing to deal with tough problems. For example, Spark Streaming is used in (Nuesch et al., 2014) to detect anomalies in water distribution networks in real time, and Apache Spark in (Domoney et al., 2015) as the tool for autonomous monitoring of city's water turbines and for automated leak detection.

Alert System for Water Quality Support

In European Union, a remote sensing tool for monitoring water quality was implemented for waters in the Mediterranean Lakes. The sensors were designed to detect cyanobacterial and other toxic substances. The system generates surveillance maps after analyzes data with the main objective to report any alerts. The generated maps are improving the MERIS and CHRIS data (from Earth observation) that were developed at the beginning of 2000 representing images from satellites focused on spectral, spatial and temporal resolutions.

Another example can be the alert system implemented on Orbigo River in Spain whose main purpose is to warn about possible droughts and prevent them to happen (Paredes-Arquiola et al., 2013). Drought planning requires preliminary identification and analysis of the risks. To reduce dryness risk, people had to understand first the climatology and make an analysis to determine the vulnerability and what people and sectors will be most affected, why these changes occur and if these relationships are changing over time. In case of Orbigo River, the demand of the system is larger than the amount of the resources available, the possibility of draughts is high. Reservoirs were constructed to maintain flood prevention and lamination during rainy seasons in autumn and spring. The reservoirs are empty before summer and full again for irrigation season. In 1998-1989, irrigation was delayed to a second plane to ensure urban water supply. The system implemented is formed by a series of piezo metric levels, streamflow, reservoir inflows and precipitation. The values taken by indicators define the drought status. For this river, there were established for levels of emergency: normality, pre- alert, alert and emergency (Haro et al., 2014).

China, the country with the most people on Earth, has developed a system named DEWS that controls the parameters of urban water quality. DEWS have a web service and provide users with water quality monitoring functions. The system is guided by control theory and risk assessment as applied to the feedback control of urban water supply systems (Lu et al., 2008).

Web Application

We developed a web application accessible via internet on every browser anytime. The benefit offered by a web application it the large scalability. Almost every person who has access to a laptop or other device is just a click away from information.

First, the web application will be implemented just for the water resources (rivers, lakes, natural pools, etc.) in Romania. On the main screen of the application there will be some menus that will include: statistics, charts, top clean/dirty water resources, search option and a history when we can find all the previous stats about that the resource.

The users can search for a water source to access more information, they will have the options to generate diagrams, to make comparisons between - for example - last year on February to same month this year. They also can generate reports for multiple water resources and will have the option to download them.

In addition to this we will place a WQI (Water Quality Index) calculator integrated on main platform where users can introduce data themselves and see the results immediately. Data feeds it's the most important thing in this project. Because of the lack of feeds, I'm obligated to divide the map into two parts: real time map (alerts generated today based on data from today) and warning map (alerts generated on last update for that water resource for example last 2 weeks). In the most cases the data will be backdated and not in real time.

For this application, we will use the most important parameters from WQI to generate alerts enumerated below: arsenic, biochemical oxygen demand, cadmium, cyanide, dissolved oxygen, fluoride, mercury, nitrate-nitrogen, pH, selenium, total coliform, and total phosphorus.

Based on legend below (see Table 1) the system will make decisions about the warnings who will be shown to users. Note that all the parameters are taking into the account to generate the warnings.

Table 1: Range – Quality semantic

Range	**Quality**
90-100	Excellent
70-90	Good
50-70	Medium
25-50	Bad
0-25	Very Bad

In Figure 2 we present the logical flow of web application scheme.

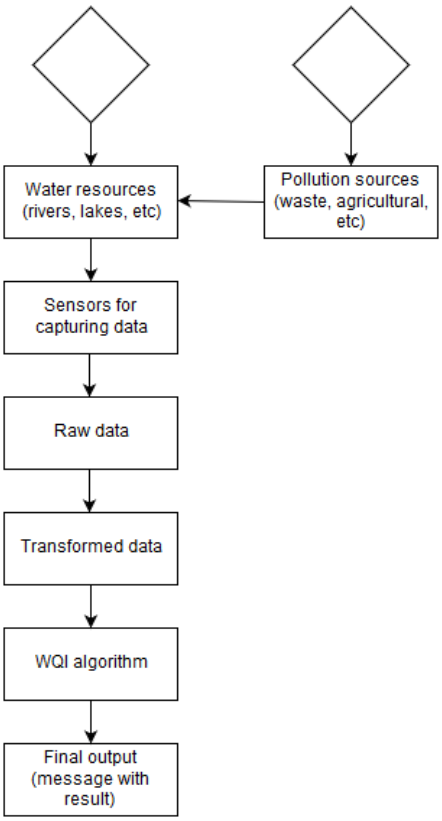

Figure 2: Web Application Scheme

Online Water Quality Monitoring (OWQM)

OWQM utilizes real-time water quality data collected from monitoring stations deployed at strategic locations in a distribution system. The data generated are continuously analyzed to support operation at system level and capture water anomalies. OWQM gives valuable understanding into real-time conditions in a distribution system. This information allows sensors to detect unusual water quality which can generate earlier, and more effective, corrections if necessary. It has other roles such as optimizing the system. It has four significant elements:

- *Data generation* which determines the water quality data produced trough OWQM. It is defined by the following decisions.

- What to monitor: the parameters monitored in the distribution system the information available to utility and the possible water incidents. Also, monitoring includes: conventional parameters: pH, specific conductance, turbidity, potential and temperature; advanced parameters: examples (TOC and UV-Vis); hydraulic parameters: pressure, flow;
- How to monitor: The sensor used for monitoring chosen parameter(s), equipment required can dramatically impact the capital and operating costs, data accuracy;
- Where to monitor: Monitoring stations can be located anywhere but should be placed in a distribution system, and can include pump stations or storage tanks (USEPA, 2015a).

- *Data communication* requires sending of OWQM data to a central storage location. Methods of communication may include digital subscriber lines, cellular networks, radio. The type and quantity of data produced, existing communication capabilities and the locations from which data must be transmitted can impact selection of data communication solution(s).
- *Information management and analysis*: receive information, processes and stores it, and make it available to users.
- *Alert investigation*: When an alert is received, utility personnel follow defined alert investigation procedures to identify its cause. In many cases, a simple review of information is sufficient to determine that an alert does not indicate anomalous water quality, and is therefore invalid. The most common causes that may occur of invalid alerts are the malfunctions of sensors and data transmission failure. If a problem can't be identified through data review, usually manual investigation is conducted at the monitoring location that induce the alert to check if accurate data is being generated and correctly communicated. Usually other samples are collected to further investigation (USEPA, 2015b).

If it is proved that an alert was caused by a water quality incident, it will be necessary to correct that with actions that mitigate potential consequences. For example, if the alert was a reaction of low disinfectant residual data, steps may be taken to increase concentrations in the area. However, if the source of the problem could not be determined, investigations will be made to the system because it can be contaminated. Standard procedures will be used based on contamination level (see Table 2).

Table 2: Goals and performances

Design Goal	**Description**
Detect water quality incidents	OWQM data can be used to detect unusual water quality conditions in distribution systems. This can contain regular system occurrence such as nitrification, rusty, turbid water. It also brings the ability to detect other substances in distribution systems resulting from pipes, negligent cross-connections, and other events, chemical spill treatment and intentional contamination.
Optimize system operation	Knowledge of a real-time water quality and improved understanding of the impact of operational changes on water quality and flow paths can improve staff to manage treatment chemicals better informing pump, valve and tank operation.
Support compliance with water quality goals and regulations	Information collected during a distribution system, particularly in areas of concern, can identify when quality goals aren't met and providing time for actions to correct potential compliance issues.
Enhance asset management	Regular data overview can reveal changes in system conditions that can affect the performance and longevity of assets such as pipes, pumps etc.

Universal Water Quality Index (UWQI)

A new upgraded index called Universal Water Quality Index was the results of the developments above. It is more simple and better to understand by 3rd party people and it main purpose it is to describe the quality of the surface water used for drinking water supply. The main addition to this index reflects the specific use for drinking water supplies rather than general supply. The UWQI is based on European Union set by Council of the European Communities in 1991 (75/440/EEC). This legislation classifies water drinking into three groups. Every group would have a different level of treatment.

- Class I: Requires basic physical treatment and disinfection;
- Class II: Requires normal physical treatment, chemical treatment and disinfection;
- Class III: Requires high physical and chemical treatment, extended treatment and disinfection.

The UWQI index will be calculated based on sub-indexes that are represented by functions which transform units and dimensions of water qualities into a variable to be represented into a common scale. The values and ranges for every parameter were calculated by water experts after elaborated studies. If the content of a sub-index is lower than the value set for class I, the value is set automatically to '100'. If the content of a sub- index is greater that the value set for class III, the value is set automatically to '0'. '50' represent the acceptable sub-index for class II. All the mathematical expression where fit for each parameter to obtain exactly these three values of '0','50' and '100' (Philadephia, 2013). The overall index formula is calculating as a sum of sub-index parameter I_i, each sub-index being multiplied by a weight w_i.

Data Evaluation and Results

Data sets proposed for tests were randomized accordingly to minimum and maximum potential values for every parameter. These are the values that I used for data in simulation assigned in application to Danube (Dunărea) River and they are not representing the real world. Data used is just to demonstrate the formulas and how the data is manipulated inside the whole system. In the image below there are exposed an example of data sets for about 14 batches of arsenic parameter. There can be found via water sub-menu by selecting the water source, in this case Dunărea River. Every batch means a complete data set of all the twelve parameters captured who are generating a full-index. Also for every line there is saved the data when was captured to keep a good track of the records. An example is presented in Table 3.

For every water source, there are also generated reports, below are some descriptive statistics that show the evaluation of data showed in Table 4. We have stats for Danube River, including every parameter, number of samples taken, mean who is the common average, median represents the middle value of samples, mode is the most common range is the difference between maximum and minimum columns.

Table 3: Example of water parameter data (Arsenic)

Water parameter data

Batch ID	Value	Date
Parametru: Arsenic (Continued on the next page)		
1	0.00000	06.06.2016
2	0.00000	07.06.2016
3	0.01000	08.06.2016
4	0.00000	09.06.2016
5	0.01000	10.06.2016
6	0.01000	11.06.2016
7	0.00000	12.06.2016
8	0.01000	13.06.2016
9	0.00000	14.06.2016
10	0.01000	15.06.2016
11	0.01000	16.06.2016
12	0.00000	17.06.2016
13	0.01000	18.06.2016
14	0.01000	19.06.2016

Page 1 of 5 (62 items) 1 2 3 4 5

Table 4: Descriptive statistics (example for Danube River)

Descriptive Statistics

Parameter	Number Of Data	Mean	Median	Mode	Range	Minimum	Maximum
Arsenic	50	0.006400	0.0100000	0.00000	0.02000	0.00000	0.02000
BOD	50	3.914000	1.7550000	5.00000	9.60000	0.10000	9.70000
Cadmium	50	0.021600	0.0150000	0.03000	0.05000	0.00000	0.05000
Cyanide	50	0.009000	0.0100000	0.01000	0.02000	0.00000	0.02000
Dissolved oxygen	50	7.382000	3.8550000	8.00000	8.90000	1.60000	10.50000
Fluoride	50	1.252000	0.6050000	1.10000	2.90000	0.00000	2.90000
Mercury	50	0.001140	0.0055000	0.00100	0.00300	0.00000	0.00300
Nitrate	50	9.760000	4.6550000	5.60000	21.60000	0.40000	22.00000
pH	50	6.116000	3.5050000	2.40000	12.60000	0.00000	12.60000
Selenium	50	0.009620	0.0100000	0.01000	0.03000	0.00000	0.03000
Total coliform	50	4682.878000	2761.4550000	45.00000	9948.90000	23.00000	9971.90000
Total phosphorus	50	0.005400	0.0100000	0.01000	0.01000	0.00000	0.01000

After few studies, it is believed that water quality assessment is far better than comparing the same data with experimentally obtained data with from existing guidelines. New indexes values are more precise for a decision to make reporting the quality of water in time and space easing the decisions to determine the maxim acceptability for each set of parameter referring to the range set in in the descriptive statistics. In Figure 3 we can see the evolution of indexes from all the 50 batches taken. In the first part, we observe a stability around 80, next its fluctuating from high to low values. All these values and indexes are automatically calculated and they are keeping updating with new data coming. We also extend the functionality to manually calculate the water quality index like Figure 4.

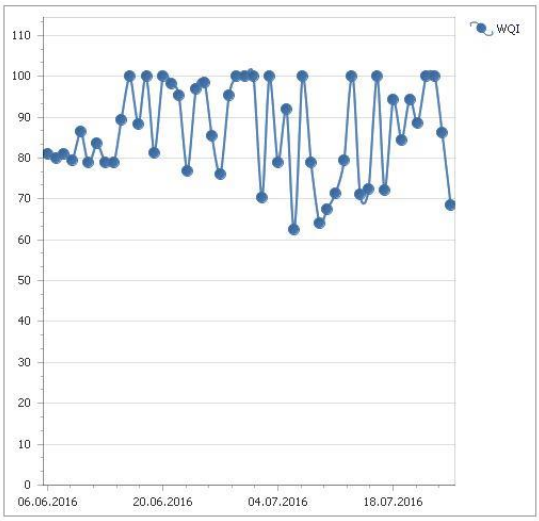

Figure 3: Evolution of Water Quality Index (example for Danube River)

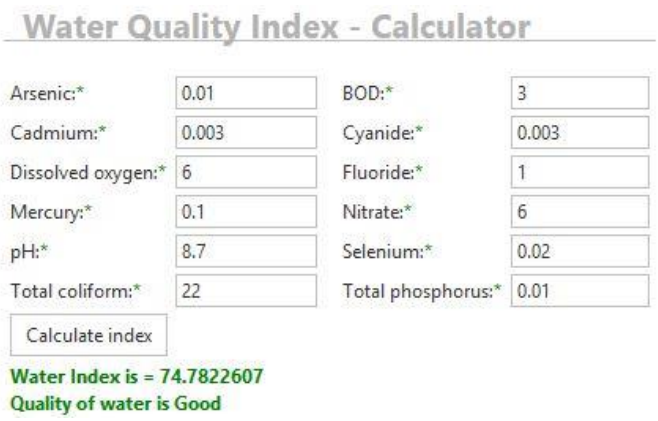

Figure 4: Water Quality Index – Calculator (example)

In Figure 5 we observe that the most influential parameters who affect the final WQI results are Selenium from Class III, pH and DO (Dissolved oxygen) from class II. To increase the WQI we need to change the parameter values to a high class. By decreasing Selenium with 0.01, pH with 1.7 and increasing Dissolved oxygen with 2 we end up with Excellent water quality with WQI at almost 94. The margins between data are so small but they are exponentially deciding the result.

Figure 5: Water Quality Index – Calculator (increased index)

Conclusion

In the first part of this paper we presented the main aspect of how ICT models can be applied in data processing as support for extracting valuable information from collected data. Several tools for water resource management are presented. The we describe the implementation of a monitoring tool for the water quality on both rivers and lakes. The part of how the data is being captured was just mentioned so we can understand how the flow works. We implement the Universal Water Quality Index, which is stronger than other classical indexes and is independently use from other research and obtained data laboratory existing guidelines purporting to improve the results based on historical data. In other words, the more data captured and covered the more precise in time is the range of the parameters data captured being able to determine the ranges of concentrations for every class.

We can conclude that data processing related research directions that need strong ICT support are very demanding in our days, considering the variety and complexity of the research field, and the necessity of targeted, specialized research teams, able to deal with different perspectives, but with deep expertise in one of them.

Acknowledgements

The research presented in this paper is supported by project Data4Water, H2020-TWINN-2015 ID. 690900. We would like to thank the reviewers for their time and expertise, constructive comments and valuable insight.

References

Boyacioglu, H. (2007). Development of a water quality index based on a European classification scheme. *Water Sa*, 33(1), 101-106.

Camp, D., Garth, C., Childs, H., Pugmire, D., & Joy, K. (2011). Streamline integration using MPI-hybrid parallelism on a large multicore architecture. *IEEE Transactions on Visualization and Computer Graphics*, *17*(11), 1702-1713.

Campbell, A.T., Eisenman, S. B., Lane, N.D., Miluzzo, E., & Peterson, R. A. (2006). People-centric urban sensing. In *Proceedings of the 2nd annual international workshop on Wireless internet* (p. 18). ACM.

Canfield Jr, D.E., Brown, C.D., Bachmann, R.W., & Hoyer, M. V. (2002). Volunteer lake monitoring: testing the reliability of data collected by the Florida LAKEWATCH program. *Lake and Reservoir Management*, *18*(1), 1-9.

Casteletti, A., Nardini, A. & Soncini-Sessa, R. (2007), Making Decisions: A Difficult Problem, In: Integrated and Participatory Water Resources Management, Soncini-Sessa R., Casteletti A., Weber E. (Eds.), 1A, Elsevier, Amsterdam, 3-36.

Cooper, C., Dickinson, J., Phillips, T., & Bonney, R. (2007). Citizen science as a tool for conservation in residential ecosystems. *Ecology and Society*, *12*(2).

de Zwart, D. (1995). Monitoring water quality in the future. Volume 3: Biomonitoring.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Dobre, C., & Xhafa, F. (2014). Parallel programming paradigms and frameworks in big data era. *International Journal of Parallel Programming*, *42*(5), 710-738.

Domoney, W.F., Ramli, N., Alarefi, S., & Walker, S.D. (2015). Smart city solutions to water management using self-powered, low-cost, water sensors and apache spark data aggregation. In *Renewable and Sustainable Energy Conference (IRSEC), 2015 3rd International.* IEEE. 1-4.

EC - Commission of the European Communities, 2002 Guidance on Public Participation in Relation to the Water Framework Directive - Active Involvement, Consultation and Public Access to Information, Common Implementation Strategy Working Group 2.9, Brussels

EC - Council of the European Communities, 2000 Directive of the European Parliament and of the Council Establishing a Framework for Community Action in the Field of Water Policy: Joint Text Approved by the Conciliation Committee 0067(COD) C5-0347/00

Ekanayake, J., Pallickara, S., & Fox, G. (2008). Mapreduce for data intensive scientific analyses. In *eScience, 2008. eScience'08. IEEE Fourth International Conference on*. IEEE.

Facchi, A., Ortuani, B., Maggi, D., & Gandolfi, C. (2004). Coupled SVAT–groundwater model for water resources simulation in irrigated alluvial plains. *Environmental modelling & software*, *19*(11), 1053-1063.

Fox, G., Bae, S. H., Ekanayake, J., Qiu, X., & Yuan, H. (2009). Parallel data mining from multicore to cloudy grids. In *High Performance Computing Workshop*. 18, 311-341.

Friess, P. (2013). *Internet of things: converging technologies for smart environments and integrated ecosystems*. River Publishers.

García, M.C.R., García, C.E.R., Brown, S., & Cordero, E. (2008). Water resource research and education in mountain communities. *Mountain Research and Development*, *28*(3), 196-200.

Geiss, F., Del Bino, G., Blech, G., N∅rager, O., Orthmann, E., Mosselmans, G., ... & Town, W.G. (1992). The EINECS Inventory of existing chemical substances on the EC market. *Toxicological & Environmental Chemistry*, *37*(1-2), 21-33.

Ghemawat, S., Gobioff, H., & Leung, S.T. (2003). The Google file system. In *ACM SIGOPS operating systems review*. ACM, 37(5), 29-43.

GWP. (2000). Integrated Water Resources Management, *TAC Background Papers*, Vol 4, Global Water Partnership.

Haro, D., Solera, A., Paredes, J., & Andreu, J. (2014). Methodology for drought risk assessment in within-year regulated reservoir systems. Application to the Orbigo River system (Spain). *Water resources management*, *28*(11), 3801-3814.

Hughes, J.D., & Liu, J. (2008). MIKE SHE: software for integrated surface water/ground water modeling. *Ground Water*, *46*(6), 797-802.

Ireson, A., Makropoulos, C., & Maksimovic, C. (2006). Water resources modelling under data scarcity: coupling MIKE BASIN and ASM groundwater model. *Water Resources Management*, *20*(4), 567-590.

Isard, M., Budiu, M., Yu, Y., Birrell, A., & Fetterly, D. (2007). Dryad: distributed data-parallel programs from sequential building blocks. In *ACM SIGOPS operating systems review*. ACM. 41(3), 59-72.

Jach, T., Magiera, E., & Froelich, W. (2015). Application of HADOOP to store and process big data gathered from an urban water distribution system. *Procedia Engineering*, *119*, 1375-1380.

Jain, M. K., Kothyari, U.C., & Raju, K.G.R. (2004). A GIS based distributed rainfall–runoff model. *Journal of Hydrology*, *299*(1), 107-135.

Khelifa, B., Amel, D., Amel, B., Mohamed, C., & Tarek, B. (2015). Smart irrigation using internet of things. In *Future Generation Communication Technology (FGCT), 2015 Fourth International Conference on*. IEEE. 1-6.

Kremlis, G., & Dusik, J. (2005). The challenge of the implementation of the environmental acquis communautaire in the new Member States. In *Seventh International Conference on Environmental Compliance and Enforcement*. Marrakech, Marocco. 9-15.

Lu, G., Wu, Z., Wen, L., Lin, C.A., Zhang, J., & Yang, Y. (2008). Real-time flood forecast and flood alert map over the Huaihe River Basin in China using a coupled hydro-meteorological modeling system. *Science in China Series E: Technological Sciences*, *51*(7), 1049-1063.

Mazumder, S. (2016). Big Data Tools and Platforms. In *Big Data Concepts, Theories, and Applications* (pp. 29-128). Springer International Publishing.

Mohn, R.A. (2006). Waterkeeper Alliance v. EPA: A Demonstration in Regulating the Regulators. *Great Plains Nat. Resources J.*, *10*, 17.

Munafo, M., Cecchi, G., Baiocco, F., & Mancini, L. (2005). River pollution from non-point sources: a new simplified method of assessment. *Journal of Environmental Management*, *77*(2), 93-98.

Nuesch, S. (2014). Real-Time Anomaly Detection in Water Distribution Networks using Spark Streaming.

O'Callaghan, J.R. (1995). NELUP: an introduction. Journal of Environmental Planning and Management 38(1), 5-20.

Odendaal, P.E. (2002). Integrated water resources management (IWRM), with special reference to sustainable urban water management. In *CEMSA 2002 Conference, Johannesburg, South Africa*.

Olston, C., Reed, B., Srivastava, U., Kumar, R., & Tomkins, A. (2008, June). Pig latin: a not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM. 1099-1110.

Paredes-Arquiola, J., Martinez-Capel, F., Solera, A., & Aguilella, V. (2013). Implementing environmental flows in complex water resources systems–case study: the Duero river basin, Spain. *River Research and Applications*, *29*(4), 451-468.

Paulos, E., Honicky, R., & Hooker, B. (2008). Citizen science: Enabling participatory urbanism. *Urban Informatics: Community Integration and Implementation*.

Refsgaard, J.C. & Storm, B. (1995). MIKE SHE. In Computer Models of Watershed Hydrology; Singh, V.P., Ed.; Water Resources Publications: Highlands Ranch, CO, USA. 809–846.

Restrepo, I. (2005). Agua y erradicacio´n de la pobreza. In: V Congreso Nacional de Cuencas Hidrogra´fica. (Abril 25–27 del 2005:Cali) CD-ROM presentaciones Congreso. Cinara.

Teodosiu, C., Barjoveanu, G., & Vinke-de Kruijf, J. (2013). Public participation in water resources management in Romania: issues, expectations and actual involvement. *Environmental engineering and management journal*, *12*(5), 1051-1063.

Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Zhang, N., ... & Murthy, R. (2010, March). Hive-a petabyte scale data warehouse using hadoop. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*.IEEE. 996-1005.

United States Environmental Protection Agency (2015). Online Water Quality Monitoring Primer - For Water Quality Surveillance and Response Systems, https://www.epa.gov/.

United States Environmental Protection Agency (2015). Water Quality Surveillance and Response Systems for Distribution System Monitoring and Management, https://www.epa.gov/.

Van Der Hoek, W., Konradsen, F., & Jehangir, W.A. (1999). Domestic use of irrigation water: health hazard or opportunity?. *International Journal of Water Resources Development*, *15*(1-2), 107-119.

Vertessy, R., O'Loughlin, E., Beverly, E., & Butt, T. (1994). Australian experiences with the CSIRO Topog model in land and water resources management. In *Proceedings of UNESCO International Symposium on Water Resources Planning in a Changing World, Karlsruhe, Germany*. 3, 135-144.

Voinov, A., & Gaddis, E.J.B. (2008). Lessons for successful participatory watershed modeling: a perspective from modeling practitioners. *Ecological modelling*, *216*(2), 197-207.

Vossen, P., Meyer-Roux, J. (1995). Crop monitoring and yield forecasting activities of the MARS project. In: King, D., Jones, R.J.A., Thomasson, A.J. (Eds.), European Land Information Systems for Agro-environmental Monitoring, EUR 16232 EN. Office for Official Publications of the European Communities, Luxembourg. 11–29.

Yu, Y., Isard, M., Fetterly, D., Budiu, M., Erlingsson, Ú., Gunda, P.K., & Currey, J. (2008). DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language. In *OSDI*. 8, 1-14.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. *HotCloud, 10,7*.

Zaharia, M., Konwinski, A., Joseph, A.D., Katz, R.H., & Stoica, I. (2008). Improving MapReduce performance in heterogeneous environments. In *Osdi*. 8(4), 7.

Zhang, D., Chen, X., & Yao, H. (2015). Development of a prototype web-based decision support system for watershed management. *Water*, *7*(2), 780-793.