IST-Africa 2021 Conference Proceedings
Miriam Cunningham and Paul Cunningham (Eds)
IST-Africa Institute and IIMC, 2021
ISBN: 978-1-905824-66-3

Big Data Analytics Capacity Development with Radio Astronomy

Nikhita MADHANPALL[1], Carolina ODMAN[2], Bonita De SWARDT[3], Vanessa McBRIDE[1], Jeremy SMITH[2], David AIKEMA[2], Kevin GOVENDER[1], Mattia VACCARI[2], Zara RANDRIAMANAKOTO [4], Tshiamo MOTSHEGWA[5], Meenakshi DEVI[6], Habatwa MWEENE [7], Nchimunya MWIINGA[7], Emmanuel NGONGA[7], Jones CHILUFYA[7], Kingsley AHENKORA-DUODU[8], Linzi STIRRUP[9], Anna SCAIFE[9]

[1]*Office of Astronomy for Development, 1 Observatory Rd, Observatory, Cape Town, 7925, South Africa*
Tel: +27 21 460 6297, Email: nikhita@astro4dev.org
[2]*Inter-University Institute for Data Intensive Astronomy, Robert Sobukwe Rd, Bellville 7535, South Africa*
Tel: +27 21 650 5273, Email: carolina@idia.ac.za
[3]*South African Radio Astronomy Observatory, 2 Fir Street, Black River Park, Observatory, 7925, South Africa*
Tel: +27 21 506 7300, Email: bonita@ska.ac.za
[4]*South African Astronomical Observatory,, 1 Observatory Rd, Observatory, Cape Town, 7925, South Africa*
Tel: +27 21 447 0025, Email: zara@saao.ac.za
[5]*University of Botswana, Private Bag UB 0022, Gaborone, Botswana*
Tel: +267 355 0000, Email: motshegwat@ub.ac.bw
[6]*Sol Plaatje University; South Africa - 10 Jan Smuts Blvd, Civic Centre, Kimberley, 8300, South Africa*
Tel: +27 53 491 0000, Email: meenakshi.devi@spu.ac.za
[7]*University of Zambia; Zambia - University of Zambia, PO Box 32379 Great East Road Campus, Zambia*
Tel: +26 021 129 5220, Email: habatwa@unza.zm
[8]*University of Leeds, Woodhouse, Leeds LS2 9JT, UK*
Tel:+44 113 243 1751, Email: kingsley.ahenkora@spacegeneration.org
[9]*University of Manchester, Oxford Rd, Manchester, M13 9PL, UK*
Tel: +44 161 306 6000, Email: linzi.stirrup@manchester.ac.uk

Abstract: To allow for African participation in the technological advances of the fourth industrial revolution and data-intensive research, there is an urgent need for data science skills capacity development on the continent. The Square Kilometre Array (SKA) Telescope in Africa will require new and innovative big data solutions and relevant technical skills, which has resulted in a drive for building capacity in this area. We present data science skills development models based on Big Data research schools and hackathons that have been run in different contexts across Africa. We describe the potential impact of these events and how this is relevant to Africa's development priorities. We present the African cloud computing facility on which events are run and findings from participant feedback. We find the events implemented by the project to be both effective and versatile, as well as adaptable to many environments, including fully virtual.

Keywords: data science, cloud computing, skills development, capacity building.

1. Introduction

The Square Kilometre Array, when completed, will be the largest scientific facility in the world and is acknowledged as one of the biggest big data machines globally. This significant multilateral investment in big data infrastructure on the African continent has been highlighted as a major driver for economic development by the countries of the SKA Africa partnership [1]. However, a lack of coordinated tertiary STEM education across Africa has been identified as a major hurdle to Africa being economically successful in the fourth industrial revolution. A skills revolution, prioritizing science, technology and innovation is one of the key goals of the African Union Agenda 2063 [2]. The Southern Africa Development Community (SADC) have proposed that reorientation of educational priorities is required to better respond to the emerging and future labour needs of 4IR, reflected in a call for the graduation of an increased number of doctorates from African universities in data-intensive fields [3]. Africa also has the richest diversity of any region of the world in terms of human genetics [4, 5]. Similarly, African biodiversity is a resource of great interest to the pharmaceutical industry as many indigenous plants on the continent may hold the key to new medication [6, 8]. Therapeutic uses of plants in Africa are often encoded in indigenous knowledge systems that are not protected by instruments applicable to intellectual property [7]. Moreover, ecosystems around the world suffer from climate change, which endangers endemic species [9]. Africa recognises the global scramble for its scientific data and there is therefore an urgent need to develop data science capacity within Africa.

In this paper, we describe a big data analytics capacity development programme that includes research schools and hackathons in a number of countries, adapting the format of each event to match specific agendas. We first describe the objectives of the programme. We then explain the methods of addressing this problem, namely data science schools and big data hackathons. We then describe the technology platform that enables those two methods. We explain that it is situated on the African continent and is, in fact, mainly used as research infrastructure. Thus, a research infrastructure can be used for broader skills development as well. We describe our results and highlight some of the feedback we have received from participants in the programme and the future perspective that this programme offers, including considerations of reproducibility. We conclude with a summary of the findings.

2. Objectives

The purpose of this work is to highlight the need for strong African data science skills and to showcase how the DARA Big Data project has worked to develop a self-sustaining research community across SKA African partnership countries. The big data challenges of the SKA Telescope led DARA Big Dara to use three focus areas to build and translate data-intensive skills, namely astronomy, agriculture and health. We present models of Big Data hackathon events designed to provide exposure to various techniques used in the field of data science and machine learning by having participants work on interesting and relevant projects that have real-world applications. We also present the goals and outcomes of events such as the Big Data Africa schools, where participants gain a broader understanding of data science and its applications beyond research. Participants also gain exposure to industry actors and “soft” skills training such as translating academic attributes to industry skills and learning about differences between academic and corporate culture. We present example implementations of the above events and the respective big data projects developed, as well as an analysis of participant feedback which provides insight into the impact of these events. Finally, we discuss the development plans of the project, highlighting the goals and vision going forward.

3. Methodology

3.1 Big Data Africa Schools

The Big Data Africa schools target students having a range of science and engineering degrees from an undergraduate to a Masters level of study, with the inclusion of a few Doctoral candidates. A large component of the school is focused on instilling good big data research practices into students. The school is held over 10 days and features a mixture of formal lectures on data science and related topics, intensive project work, and industry and career skills development. The school teaches important techniques for working with large datasets and focuses on group learning and working with 'real-life' big data, in addition to networking with peers and key industry contacts. The general format of the Big Data Africa schools consists of the three main learning components:

1. Formal lectures aim to introduce students to basic concepts in data science, and to familiarise them with the different tools and techniques needed for working with large data sets;
2. Project work where students apply the theory that they have learnt to a practical project. These projects are prepared ahead of the school and cover several disciplines. The practical project is chosen by the student at the beginning of the school;
3. Industry skills development. This includes guest presentations and workshop sessions with the aim to expose students to the broader applications of data science and Big Data beyond academia as well as professional aspects of working in the private sector.

The lectures at the Big Data Africa schools introduce cross-disciplinary data science topics that can be applied to any science research project involving large data sets. Lectures include introductory-level statistics, Python programming and data visualization for data science, as well as a module on machine learning. Informal hands-on practical sessions take place between lectures where students work on the pre-selected projects in smaller groups. Group work enhances the students' learning by focusing on peer-to-peer learning while working on a "real world" problem. This teaches the students how to build relationships within a team and amongst their peers, which is critical for their development as young researchers. Each group is also supported by a tutor, a slightly more senior researcher who is an expert on the subject. The tutors help the students overcome practical problems so that they can focus on learning data science. The projects and their data are placed on a cloud computing platform where virtual machines are pre-configured for each project.

Table 1: Big Data Africa research projects

Example Projects	*Description*
Radio astronomy	Detection of radio frequency interference in radio astronomy data The Search for Extraterrestrial Intelligence (SETI) in radio astronomy data Hunting down Fast Radio Bursts with Breakthrough Listen
Health	Detection of retinopathy in medical imaging of human eyes Prediction of epileptic seizures using intracranial EEG recordings Using 3D images to find cancer detection patterns in patient data
Road safety	Analysis of traffic data and prediction for road accident prevention
Cyber security	Analysis of web traffic data and identification of malicious queries
Astronomy	Detection of planets in Kepler satellite data

Agriculture	Mapping and analysing crop phenology using Satellite Earth Observation data
Bioinformatics	Finding meaningful patterns for ovarian cancer detection in genomic patient data

Table 2: DARA Big Data Africa schools and participation

Date	*Location*	*Number of participants*	*% Female participants*
April 2017	Cape Town	38	26
May 2018	Madagascar	19	42
September 2018	Cape Town	26	46
October 2019	Cape Town	25	48

3.2 Big Data Hackathons

In addition to the Big Data Africa schools, DARA Big Data hackathons have been implemented in various SKA African partner countries since 2018. These hackathon events are shorter (2.5 days), more focused events which aim to provide exposure to data science and machine learning techniques and grow confidence in applying these techniques to solve real world problems. The current format of the hackathons is as follows: To open the event, there is an afternoon dedicated to presentations by professionals from the three partner institutions, as well as invited guest speakers, on various topics of interest such as, data science, machine learning, industry applications, data science for development, etc. On day two of the event, participants commence the hackathon by working through guided tutorials (in teams of four or five) with the assistance of ample tutors. On day three, teams tackle a short hackathon task using the skills learned from the tutorials on day two.

Table 3: Hackathon projects

Example Projects	*Description*
Social Networks	Sentiment analysis of Twitter feed about Covid-19
Product recommendation	Building a movie recommender engine using machine learning
Earth observation	Floodplain analysis using satellite imaging
Image classification	Web scraping and image classification
Music classification	Build a music classification and recommendation engine
Astronomy	Detection of pulsar signals in noisy radio astronomy data

Table 4: Hackathon events

Date	*Location*	*Number of participants*	*% Female participants*
November 2018	Botswana	39	20.5
October 2019	Namibia	24	29
March 2020	South Africa	33	24
September 2020	Zambia	22	32
November 2020	Online	34	20.5

Both hackathons and the schools end with group presentations (a key career development skill), showing the problems that were looked at, how they used the data and the conclusions they drew from it. Presentations are judged on participant development, project knowledge and the visual presentation of data. DARA Big Data postgraduate students often serve as project facilitators and leaders in both schools and hackathons, forming a pipeline of talent that will hopefully continue to extend across Africa.

4. Technology Description

The following cloud computing technology stack is used to support the schools and the hackathons. The Ilifu Cloud Computing Research Facility was created with funding from the Data Intensive Research Initiative of South Africa as a Tier 2 (regional) facility and supplemented by various institutional and user-supplied contributions including from IDIA. It is governed by a consortium of universities in the Western Cape and Northern Cape of South Africa and representatives from the national government. The equipment is installed at the University of Cape Town. The Ilifu cloud consists of several petabytes of storage and over 100 compute nodes and runs Openstack [12] as the cloud infrastructure layer. It is primarily used by astronomers and bioinformaticians for their research computing needs.

Virtual machines on the Ilifu cloud are made available for the Big Data Africa schools and hackathons to fulfil their computing needs. The virtual machines are created using a generic school/hackathon template that is then customized to meet the requirements of each event and each project, and then used to build a common base image for each of the event's virtual machines. The base template is created using Packer [14] to assemble the different software elements needed for the events from their original sources. This approach prevents hidden dependencies that might arise if changes are manually made to a static, pre-existing image and allows us to update components in the image as needed. A script has been created to generate a set of accounts for each new virtual machine and output a list of passwords that can then be distributed to the participants upon their arrival.

Users are primarily presented with a Jupyter Notebook interface [14], using JupyterHub used to manage the multiple users. The Jupyter web interface is secured by using SSL certificates freely available from Let's Encrypt [15]. In some instances, SSH access to the virtual machines might also be provided. The per-event customization of the virtual machine images means that a user can be presented upon initial login with a Jupyter notebook that they can then work with throughout the duration of the event. For simplicity, although Jupyter and JupyterHub are both written in Python, Python virtual environments are used to separate the Python environment they utilize from that used within the notebook environments. This enables a project to use a set of python packages that are incompatible with those required by the Jupyter installation without problems and has happened in at least one instance.

All projects developed for the Big Data hackathons are added to the DARA Big Data GitHub repository [16] and are publicly accessible. This repository serves as a learning tool for individuals who may not have been able to take part in a hackathon, but would like to go through the tutorials and work on the hackathon task in order to further their understanding of data science and machine learning techniques. The repository is also intended to equip institutions/groups that may be interested in running independent hackathon events with the educational resources necessary to do so.

5. Results

The Big Data Africa school has been run three times since 2018 and participants have come from diverse scientific research areas. The Schools are intensive and immersive, allowing for a faster progression of skill development. Prior programming experience is required and participants report greatly increased knowledge in technical areas such as cloud computing, image processing and machine learning. A common feedback theme is that participants greatly appreciate working with peers from across the continent, developing both communication skills and their networks.

The first Big Data hackathon was hosted in November 2018 at the University of Botswana, as part of International Data Week. Participants were a mixture of undergraduate and postgraduate students and also industry professionals. In feedback after the event 100% of students reported they developed new skills and 88% thought it tested their coding skills. A second hackathon was held at the Namibia University of Science & Technology in October 2019. Afterwards participants reported that their main motivation for taking part was learning to code to improve career prospects. Most indicated that they would use their new skills for study and work, with some hoping to pursue a PhD in related fields. Areas where they thought machine learning skills were applicable included computer security, public health, geology, astronomy and foreign exchange. Much of this reflects similar feedback from other events (illustrated in Figure 1). Overall participants found themselves gaining skills and confidence in data science and could see themselves pursuing data science as a career option.

In March 2020 a hackathon was held in South Africa at Sol Plaatje University, a young and mainly teaching-focussed institution that offers Data Science as a discipline. A key goal with this event was to raise awareness of cloud technologies for research and increase the University's participation in Ilifu, a high-performance cloud computing facility. This was achieved as five research projects from Sol Plaatje faculty members were then onboarded to the Ilifu cloud. This demonstrates that the hackathon model can be a viable means of strengthening research capacity in teaching-heavy institutions.

The outbreak of the global pandemic meant that the hackathon model had to change and adapt. An event was held at the University of Zambia in September 2020 where students were gathered in teams at the University's computer labs with two tutors physically present and three more tutors online. Talks were given using Zoom, including the student presentations at the end of the event. Unlike previous events, where participants had been mainly at the postgraduate level, these students were undergraduates with little Python experience. Nonetheless it proved to be very successful in terms of student development, as evidenced by their presentations and their positive post-event feedback. The main challenges faced were the unstable internet connection and difficulties with remote tutoring, leading to the two tutors onsite being somewhat overwhelmed with requests for help from students.

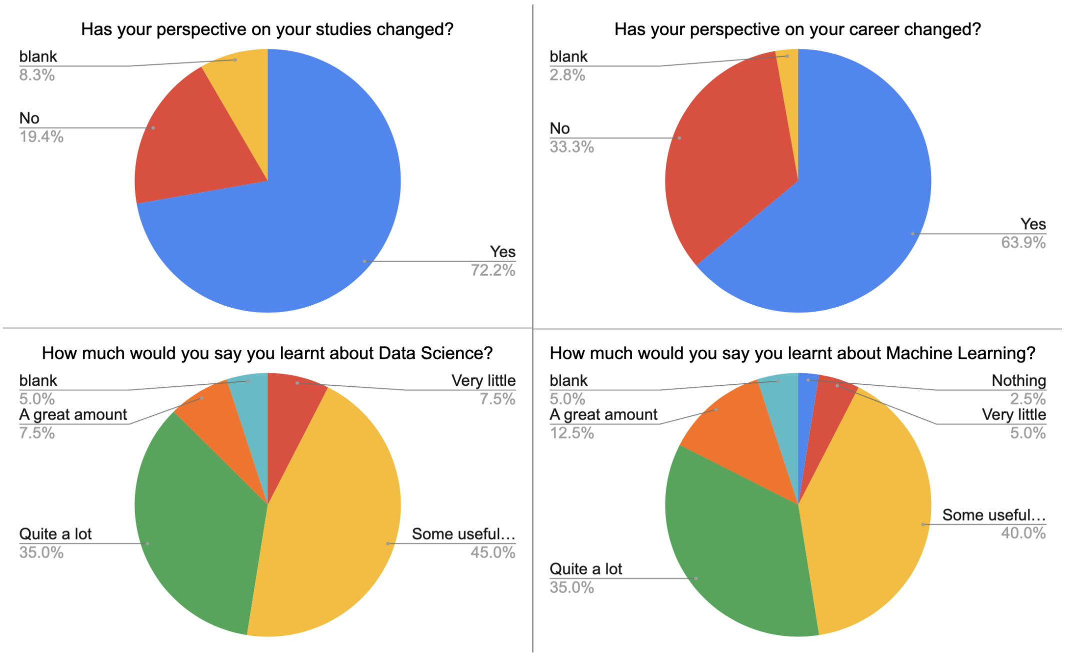

Figure 1: Extract of feedback from the University of Zambia and SGAC hackathons in 2020. (N=40). The yellow category in the second row is "Some useful things".

The latest event in the series was organised in collaboration with the African chapter of the Space Generation Advisory Council (SGAC), an organisation of young space professionals and enthusiasts. It was held in November 2020 and was fully remote with participants from 12 African countries grouped into international teams. Tutoring and presentations were all done online and overall it went very smoothly. Learning to collaborate and communicate remotely was a learning objective of the event which was met successfully as teams worked well together and produced good results.

A key focus of the events is a drive to increase female participation in data science, as they are currently well under-represented in this field in Africa (and internationally). A recurring problem is that neither the schools nor the hackathons receive as many applications from females as males, however the Big Data Africa schools have seen a significant 15% increase in female applicants since they began; at the most recent one 35% of attendees were women. Great care is taken with the schools to ensure as even a gender split as possible in participants; out of 108 participants in the 4 schools 42 have been female. The same thing is not possible with the hackathons where all applicants are accepted and the number of female applicants averages around 27% for each event. In total 154 participants have taken part in hackathons, of which just 37 have been female. It is notable that in both pre- and post-event feedback females rate their skills lower than their male peers, indicating a lack of confidence for female participants, however two hackathon events have ended with majority female teams taking the top prize. Female participants are always placed in teams with at least one other woman to reduce the risk of them feeling undermined or side-lined, however more needs to be done to attract their participation in the first place.

6. Developments

There are many other programmes across Africa seeking to develop capacity in big data analytics. The Deep Learning Indaba [17] started off as an academic community building exercise and has blossomed into a pan-African network of machine learning researchers and professionals with high impact. The Human Heredity and Health in Africa programme (H3A Bionet) has developed online training in bioinformatics with a train-the-trainers model and hybrid remote and in-person learning [18], and they are extremely successful at developing bioinformatics skills, which are also transferable, like astronomy skills. The private sector has also developed data science programmes, for example the EXPLORE Data Science programme [19], to accelerate the development of data science skills for industry in South Africa, and Zindi is an African data science challenge platform similar to the well-known Kaggle, which also contributes to growing data science skills [20].

This may seem like a crowded landscape but the need for data science skills still outsizes their availability greatly. We find that our big data schools and hackathon models are very versatile and adaptable to local and regional contexts and we are building a database of projects that can be reused by others. In the future we hope to further streamline the process, possibly setting a reliable standard for tutors, with acknowledgment of their support. We will also start networking the alumni of the schools and hackathons and hope to build a supportive group, as well as a pool of skilled African data scientists to recommend to our research communities and to industry partners.

A long-term vision of the partner institutions is to ensure that the hackathons initiative is scalable as well as sustainable. Reproducibility practices are therefore crucial when developing the hackathon resources. With the use of public resources available in the DARA Big Data GitHub repository, containerization, and common clouds such as Google Colaboratory and the Kaggle cloud, hackathons can be run independently by various groups/institutions across Africa.

7. Conclusions

In this paper we describe a model of a broad data science skills development programme set up in response to the need for Africa to be able to seize the opportunity presented by African scientific data. We have developed a model of big data analytics research schools and hackathons that we have deployed over the last 3 years across the African continent with the aim to develop strong big data skills for research using the SKA as a window of opportunity and supported by a technological research platform located on African soil. We find that data science skills development is highly relevant not only for developmental challenges, but also to equip Africa to face the fourth industrial revolution. While DARA Big Data is not the only programme in this landscape, we find that the demand from students for opportunities to develop their data science skills as well as the demand from industry far outweighs the capacity of programmes and we find that our model complements others well. Our participants and our tutors both develop technical skills as well as soft skills, being exposed to industry ways of working as well as academic research, and the model has proven to adapt very well to the world of remote work and remote collaboration that was precipitated by the COVID-19 pandemic. Going forward we aim to make this programme self-sustained by sharing our project resources and will continue to streamline the events and connect our participants to ensure a broad societal impact.

Acknowledgement

The programme, Development in Africa with Radio Astronomy (DARA) Big Data, is an international collaboration jointly funded by the UK Newton Fund and the South African Department of Science and Innovation (DSI). The implementation of the programme is led by the University of Manchester (UK) and the South African Radio Astronomy Observatory (SARAO), in collaboration with the Inter-University Institute for Data Intensive Astronomy (IDIA) and the IAU Office of Astronomy for Development (OAD).

References

[1] https://www.sarao.ac.za/science/avn/
[2] https://au.int/en/agenda2063/aspirations

[3] Statement by His Excellency Dr. Hage G. Geingob, President of the Republic of Namibia and Chairperson of SADC on the occasion of the launch of the ILO Global Commission on the Future of Work report. March 1, 2019. https://op.gov.na/documents/84084/805187/Statement+by+President+Hage+G+Geingob+at+launch+of+ILO+Report+-+Durban+2019+Final.pdf/2f9a6a23-4ed5-4998-9cf4-56cf3d84ac17?version=1.0

[4] A. Mickey, *The genetic diversity in Africa is greater than in any other region in the world.* https://blogs.bcm.edu/2018/07/19/genetic-diversity-in-africa-is-greater-than-in-any-other-region-in-the-world/

[5] Tucci, S., Akey, J.M. *The long walk to African genomics*. Genome Biol 20, 130 (2019). https://doi.org/10.1186/s13059-019-1740-1

[6] A. Gurib-Fakim, *Capitalize on African biodiversity*. Nature 548, 7 (2017) doi:10.1038/548007a

[7] Reihling, H.C. *Bioprospecting the African Renaissance: The new value of muthi in South Africa*. J Ethnobiology Ethnomedicine 4, 9 (2008). https://doi.org/10.1186/1746-4269-4-9

[8] N. Makunga. *How changes in African traditional medicine research can benefit South Africa*. The Conversation, 3 September 2015. https://theconversation.com/how-changes-in-african-traditional-medicine-research-can-benefit-south-africa-46486

[9] V. Neergheen-Bhujun et al. *Biodiversity, drug discovery, and the future of global health: Introducing the biodiversity to biomedicine consortium, a call to action*. J Glob Health. 2017 Dec; 7(2): 020304. doi:10.7189/jogh.07.020304

[10] S. Kendrew, C. Deen, N. Radziwill, S. Crawford, J. Gilbert, M. Gully-Santiago, and P. Kubánek *The first SPIE software Hack Day* Proc. SPIE 9152, Software and Cyberinfrastructure for Astronomy III, 915202 (2014). https://doi.org/10.1117/12.2075357

[11] Kendrew, S., Simpson, R. J., Lintott, C. J., Crawford, S. M., Smith, A., Ödman-Govender, C., Bauer A., Smethurst B., Nekoto, W. *Ten Years of .Astronomy: Scientific and Cultural Impact.* Bulletin of the AAS, 51(4). (2020) https://baas.aas.org/pub/2019i0203

[12] https://www.openstack.org/

[13] https://www.packer.io/

[14] https://jupyter.org/

[15] https://letsencrypt.org/

[16] https://github.com/darabigdata

[17] https://deeplearningindaba.com/

[18] https://www.h3abionet.org/training

[19] https://explore-datascience.net

[20] https://zindi.africa