Big data processing with Apache Spark in university institutions: spark streaming and machine learning algorithm

Emmanuel Boachie*

School of Computer Science and Technology,
Wuhan University of Technology,
Wuhan 430063, China
and
Department of Computer Science,
Kumasi Technical University,
Box 854, Kumasi, Ghana
Email: mmel179@hotmail.com
*Corresponding author

Chunlin Li

School of Computer Science and Technology,
Wuhan University of Technology,
Wuhan 430063, China
Email: 1957542903@qq.com

Abstract: Data processing is an effective tool for educational sector, which can improve admission selection procedures and decisions. Most research papers focus on computational and theoretical aspect of education though little effort have been put on technological aspect of applying data mining techniques on students admission process. We therefore design a simple spark streaming framework together with machine learning algorithm to guide admission processing. We implement the spark streaming model and the proposed machine learning algorithm in a selected university using its admissions' data. The focus is on the number of students that can be admitted and those that should be rejected to reduce time and cost. The case study we evaluated show the practical usefulness of Spark streaming and machine learning algorithm for data processing in a real-time to reduce time and cost. The experiment results also confirm meaningful graphical interpretation of data using spark streaming and machine learning algorithm for students' selection for admissions.

Keywords: spark streaming; big data processing; machine learning algorithm.

Reference to this paper should be made as follows: Boachie, E. and Li, C. (2019) 'Big data processing with Apache Spark in university institutions: spark streaming and machine learning algorithm', *Int. J. Continuing Engineering Education and Life-Long Learning*, Vol. 29, Nos. 1/2, pp.5–20.

Biographical notes: Emmanuel Boachie is currently a PhD student at the Wuhan University of Technology, Department of Computer Science and Technology, Wuhan, 430070, China. He received his Master's in Information Technology, Multimedia and Education at the University of Leeds, UK and

lectured at the Ghana Baptist University College from 2007 to 2011. He joined Kumasi Technical University from 2011 to 2016 as a Lecturer. His major in PhD program is big data.

Chunlin Li is a Lecturer and the Vice Dean at the Wuhan University of Technology, Department of Computer Science and Technology, Wuhan, 430070, China. She specialises in networking and cloud computing, big data and other computer science related areas.

1 Introduction

Data are increasingly growing in university institutions because of the rising number of students and courses. It is unproductive or almost unlikely to use relational databases management systems to manage this diversity of data streaming into the system. Relational databases management system can easily manage homogenous data but not all types of data are homogenous. There are other heterogeneous data in the academic institutions which includes video, audio, pictures, location data, and so on (Okur and Buyukkececi, 2014). These data type to be stored and managed are very huge. This means that, it will occupy huge space ranging from the size of petabyte (1,015) to yottabyte (1,024). The utmost accessible space for SQL 2014 for business intelligence, enterprise, standard and web applications is 524 petabytes (524 × 1,015). This indicates that, with the exponential growth of data retrieved in the tertiary sector on daily basis, it will be complex to store data with this type of RDBMS.

Information is important to the management board of every academic institution. With some basic features of processed data being accessibility (data needs to be simply available to the authorised users in that they can get it in the appropriate structure and at the appropriate moment to meet their wants.), timely (timely data is provided the time it is required) and verifiability (it shows that one can verify it to be certain that it is accurate, maybe by verifying a lot of sources for the similar information). The procedure of accessing information must also be timely (Jihong and Yue, 2017).

In university institutions, more especially developing countries, different set of information/data are derived from the various departments and other functional sections. In academic institutions information/data gathered is limited to the departments which gathered it but centralisation of the various data in the various functional sections does not exist (Jomafuvwe, 2018). This makes it difficult for the management committee to take decisions based on relevant information needed. All the departments and other functional sectors to manage their data separately and provide them to the management committee when they are asked for. These data are normally provided manually, most at times in hard copies, by the various functional sectors to the management committee. This situation normally prevents the management committee to have relevant information when the need arises for decision making.

Data processing is an effective tool for educational sector which can improve admission selection procedures and decisions. Most research papers focus on computational and theoretical aspect of education though little effort have been put on technological aspect of applying data mining techniques on students' admission selection. The objective of the study is to design cost effective approach and innovative tool to process university students admission data in real-time. The paper outlines a simple spark streaming framework and machine learning algorithm to guide admission processing and focuses on the number of students that can be admitted and rejected to enhance management in the university institutions. The experiment conducted using spark streaming and machine learning algorithm confirm the proposed tools as cost effective and innovative tool for university students' admission data. The result of the experiment can help the researcher to apply spark streaming and machine learning algorithm for data processing as it reduces time and cost and provides comprehensive graphical interpretation of data as compared to using Excel, Tableau and other statistical tools.

The rest of this paper is organised as follows: related work is discussed in Section 2. Proposed spark streaming model and proposed machine learning algorithm are discussed in Section 3. Data source is indicated in Section 4, experiment and experimental results are presented in Sections 5 and 6. Section 7 briefly discussed the experiment evaluation. Finally, conclusion and future works are provided in the last section.

2 Related work

Big data is an acquisition of knowledge system which is in advance altering the entities of knowledge and social theory in a lot of areas while also having the ability to change a theory of an administrative decision-making (Long and Siemen, 2014). Big data incorporates the emergent research area of studying analytics that is already increasing in the area of education. Nonetheless, research in learning analytics has been greatly cut down to assessing the indicators of individual students and classroom ability. Big data seeks new chances and setbacks for university institutions. Lavvitt (2017) showed that big data proposes the most dramatic model in effectively using the vast array of data and ultimately modifying the future of university institutions. The adoption of big data in university institutions was also made known by Hrabowski et al. (2011) who indicated that technological developments have ultimately served as catalysts for the move towards the advancement of analytics in university institutions.

In the area of university education, big data can connote the interpretation of a vast range of institutional and functional data mobilised processes aimed at examining the institutional performance and progress with the view of predicting future performance and indicate potential issues related to teaching and learning, academic programming and research. Others also showed that to meet the requirements of improved productivity, university institutions have to introduce the tool of analytics into the system. As a new area within the institutions, many scholars have affirmed that big data model is well structured to address some of the main setbacks now facing university institutions (Siemens, 2011). At the initial stage much of the work on analytics within university institutions is emerging from interdisciplinary research, spanning the areas of computer science and information science, educational technology, statistics and mathematics. A key component of the present work on analytics in education is focused on data mining.

Big data in university institutions also captures database systems that save great amount of longitudinal data on students' from the beginning to a particular transaction and activity on teaching and learning. When students applying learning technologies, they can leave data trails behind that can indicate their social connections, intentions, goals and sentiments. Researchers can apply such data to evaluate the trends of student ability over time from a semester to a semester or from a certain year to another year. On an advance stage, it could be defended that the added value of big data is the ability to identify important data and change it into usable information by identifying trends and deviations from trends. OECD (2013) report showed that big data is currently well structured to begin addressing some of the core setbacks now facing university institutions. The report proposed that it could be a basis on which university institutions can redevelop both their business model and present together the proof to assist in making decisions about educational outcomes.

From an organisational learning perspective, it is well acceptable that institutional efficiency and adaptation to change depends on the analysis of right data and that existing technologies help institutions to have an in depth knowledge about data with previously unachievable levels of speed, accuracy and sophistication (Raymond, 2014). As technologies continue to penetrate all facets of university institutions, important data is being extracted by students, computer applications and systems (Brower, 2017).

Moreover, big data analytics can be adopted to evaluate student's entry on a course assessment, discussion board entries, that can extract more transactions per student per course. These data would be gathered in real or almost real-time as it is transacted and then assessed to propose courses of action. As Clark (2017) showed that "[learning] analytics are a basic tool for informed modification in education" and provide proof on which to form understanding and make informed decisions.

2.1 *The value of big data in university institutions*

Big data could address the setbacks regards to finding information at the right time when data are deployed across many unlinked different data systems in institutions. By identifying procedure of aggregating data across systems, big data can improve decision-making capability (Baker and Inventado, 2016).

2.2 *Implementation setbacks*

There are many predictable setbacks associated with the execution of analytic procedures for big data in university institutions. A number of these include setbacks related with securing consumers to acknowledge big data as a channel for applying fresh processes and alter administration. Secondly, there is an astonishing charge associated with introducing mathematical formulas to extract data gathering and keeping a process which is predisposed to be time wasting and technical. Again, several organisational information schemes are not interoperable, so cumulating managerial information; classroom and online data can possibly indicate extra setbacks (Daniel and Butson, 2013).

Spectacular progress in information gathering, processing power, information communication and keeping capacities are facilitating many institutions to combine their variety of databases into data banks. In the era of plentiful information, tertiary institutions related to institutions, healthcare or government sectors have some of the similar grounds for applying analytics, particularly in the sections of financial effectiveness, enlarging local and worldwide effect, deal with innovative monetary support models throughout an altering cost-effective climate, and retorting to the requirements for better responsibility.

Within tertiary institution, information are on the rise, though the majority of it is spread out transversely desktops, departments and come in a variety of layouts, making it complex to generate or merge. To successfully access this information, the capability to examine varied information sets is required. In spite of their source, and combination data kept in silos within institutions, and administrating the information while saving insightful data across databases, is a core demand for application of big data in higher schooling.

Analytics also has the ability to assist students and teachers to be aware of risk signals before menace to studying success achieved. Nevertheless, broad institutional reception of analytics requires a clear organisational plan and the application of analytics software packages (White, 2011).

Moreover, information storage implements normal data layouts. Every section will generate outcomes which are standardised with all the other sections, resulting in additional precise information demonstration. Finally, an information storehouse can keep huge quantities of past information that can be willingly tested and to examine diverse periods of time to make prospect forecasting. Data processing is an effective tool for educational sector, which can improve admission selection procedures and decisions. Most research papers focus on computational and theoretical aspect of education though little effort have been put on technological aspect of applying data mining techniques on admission selection.

3 Spark streaming

Spark streaming is a component which enhances processing of live streams of data. Examples of data streams include log files produced by production web server, or line ups of messages including status revision posted by production web server. Spark streaming offers an API for manoeuvring data streams that intimately matches the spark core's RDD API, making it simple for programmers to study the scheme and move between applications that manoeuvre information kept in memory, on disk or arriving in genuine occasion. Beneath its API, spark streaming has been devised to offer similar level of error acceptance, throughput and scalability as spark core. Spark comes with in-build machine learning library and spark streaming which is capable to capture information in genuine occasion for analytics ('Spark Release 2.0.0', 2014).

3.1 Spark streaming model

The proposed spark streaming model is real-time data streaming. Apache Spark deals with real-time streaming of outside data.

3.1.1 Data modelling

Apache Spark comes with MLlib, for modelling set of data for predictions (Zaharia et al., 2014b). Our proposed framework used support vector machine (SVM) and regression for modelling and training dataset for identifying patterns and making effective predictions.

1 *SVM:* are a set of supervised learning methods used for classification, regression and outliers' detection. The advantages of SVM are: effective in high dimensional spaces. Still effective in cases where number of dimensions are greater than the number of samples.

Uses a subset of training points in the decision function (called support vectors), so it is also memory efficient.

- Versatile: different kernel functions can be specified for the decision function. Common kernels are provided, but it is also possible to specify custom kernels (Zaharia et al., 2014a).

Below shows the model design with the SVM.

Give $y \in \{1, -1\}^n$ n training vectors $x_i \in \mathbb{R}^p$, $i = 1, \ldots, n$, in two classes, and a vector, SVC solves the following primal problem:

$$\min_{w,b,\varsigma} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \varsigma_i$$

$$\text{subject to } y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \varsigma_i,$$

$$\varsigma_i \geq 0, i = 1, \ldots, n$$

Its dual is

$$\min_{\alpha} \frac{1}{2} \alpha^T Q \alpha - e^T \alpha$$

$$\text{subject to } y^T \alpha = 0$$

$$0 \leq \alpha_i \leq C, i = 1, \ldots, n$$

where e is the vector of all ones, $C > 0$ is the upper bound, Q is an n by n positive semi definite matrix, $Q_{ij} \equiv y_i y_j K(x_i, x_j)$, where $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is the kernel. Here training vectors are implicitly mapped into a higher (maybe infinite) dimensional space by the function ϕ.

The decision function is:

$$\operatorname{sgn}\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + \rho \right)$$

While SVM models derived from *libsvm* and *liblinear* use C as regularisation parameter, most other estimators use *alpha*. The exact equivalence between the amount of regularisation of two models depends on the exact objective function optimised by the model. For example, when the estimator used is *sklearn.linear_model.Ridge* regression, the relation between them is given as

$$C = \frac{1}{alpha}$$

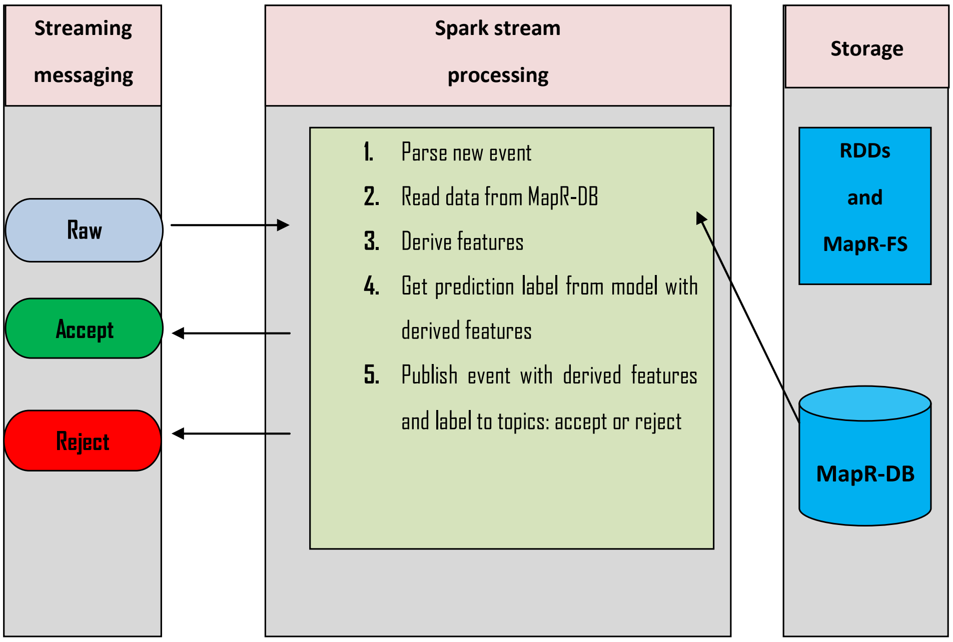
2 *Linear regression:* simple linear regression is a statistical method that allows us to summarise and study relationships between two continuous (quantitative) variables. This lesson introduces the concept and basic procedures of simple linear regression. We will also learn two measures that describe the strength of the linear association that we find in data. The proposed framework used regression to assess the GPA of students from low and high senior high school's university entrance test score.

We summarise by the formula

$$\mu Y = E(Y) = \beta 0 + \beta 1x$$

We can also express the average university entrance test score for the i^{th} student, $E(Y_i) = \beta 0 + \beta 1xi$. Of course, not every student's college entrance test score will equal the average $E(Y_i)$. There will be some error. That is, any student's response y_i will be the linear trend $\beta 0 + \beta 1xi$ plus some error $\in i$. So, another way to write the simple linear regression model is $y_i = E(Y_i) + \in i = \beta 0 + \beta 1xi + \in i$.

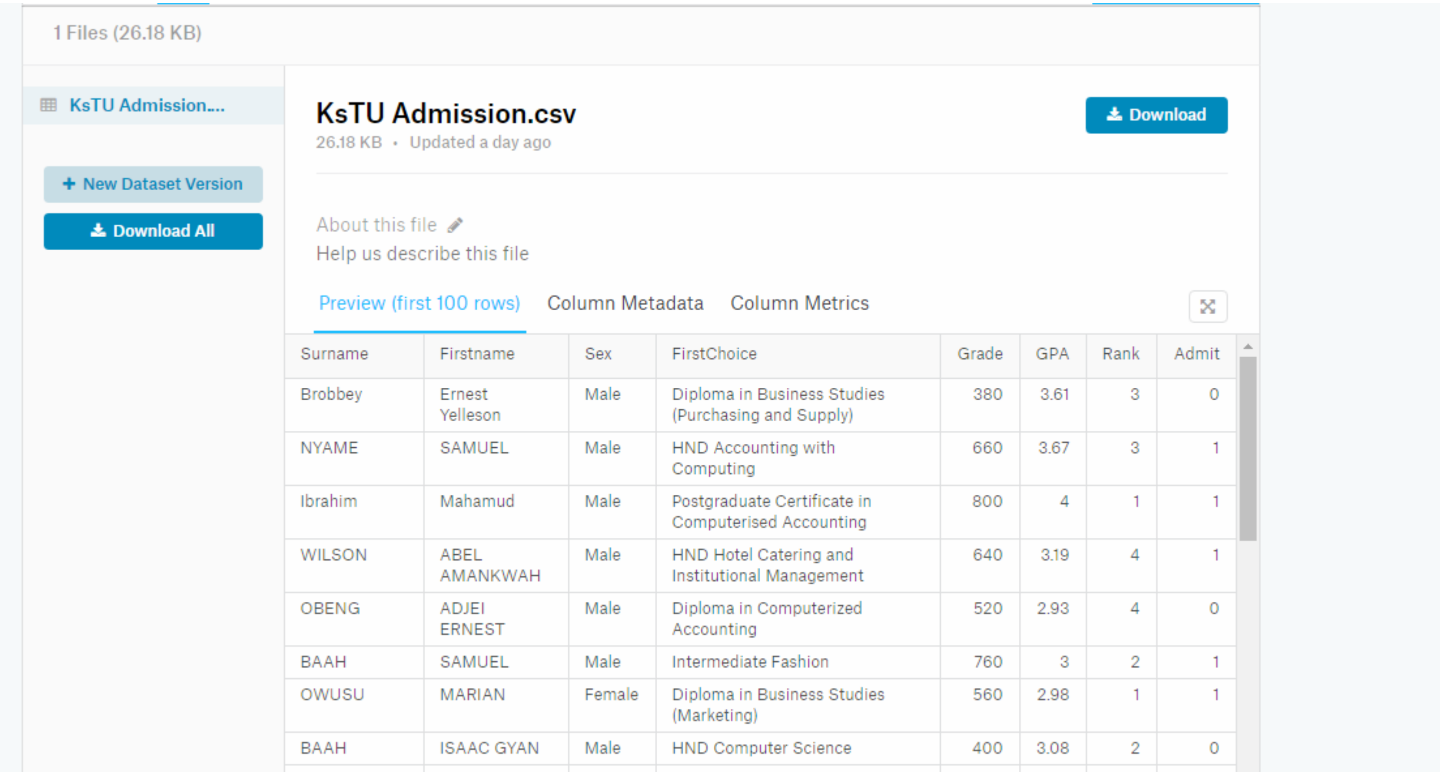
Figure 1 Spark streaming model (see online version for colours)


Streaming messaging
Spark stream processing
Storage
Raw
Accept
Reject
1. Parse new event
2. Read data from MapR-DB
3. Derive features
4. Get prediction label from model with derived features
5. Publish event with derived features and label to topics: accept or reject
RDDs and MapR-FS
MapR-DB

4 Data sources

The data selected for the experiment came from the Kumasi Technical University, Ghana for processing and analytics. The data selected for the experiment ranges from 2006 to 2016 students' admissions. The experiment applying the proposed tools will determine which students to be admitted and those not to be admitted based on their grades.

Figure 2 Admission dataset connect to databricks (see online version for colours)



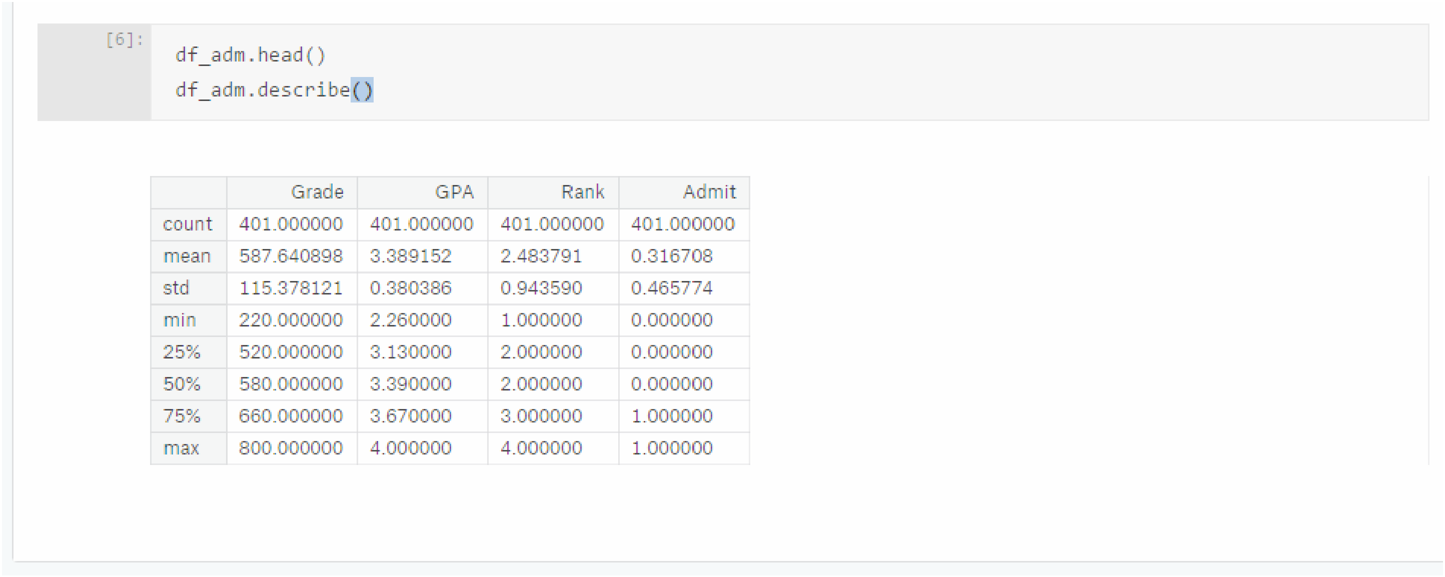
KsTU Admission.csv

26.18 KB • Updated a day ago

Help us describe this file

Surname	Firstname	Sex	FirstChoice	Grade	GPA	Rank	Admit
Brobbey	Ernest Yelleson	Male	Diploma in Business Studies (Purchasing and Supply)	380	3.61	3	0
NYAME	SAMUEL	Male	HND Accounting with Computing	660	3.67	3	1
Ibrahim	Mahamud	Male	Postgraduate Certificate in Computerised Accounting	800	4	1	1
WILSON	ABEL AMANKWAH	Male	HND Hotel Catering and Institutional Management	640	3.19	4	1
OBENG	ADJEI ERNEST	Male	Diploma in Computerized Accounting	520	2.93	4	0
BAAH	SAMUEL	Male	Intermediate Fashion	760	3	2	1
OWUSU	MARIAN	Female	Diploma in Business Studies (Marketing)	560	2.98	1	1
BAAH	ISAAC GYAN	Male	HND Computer Science	400	3.08	2	0

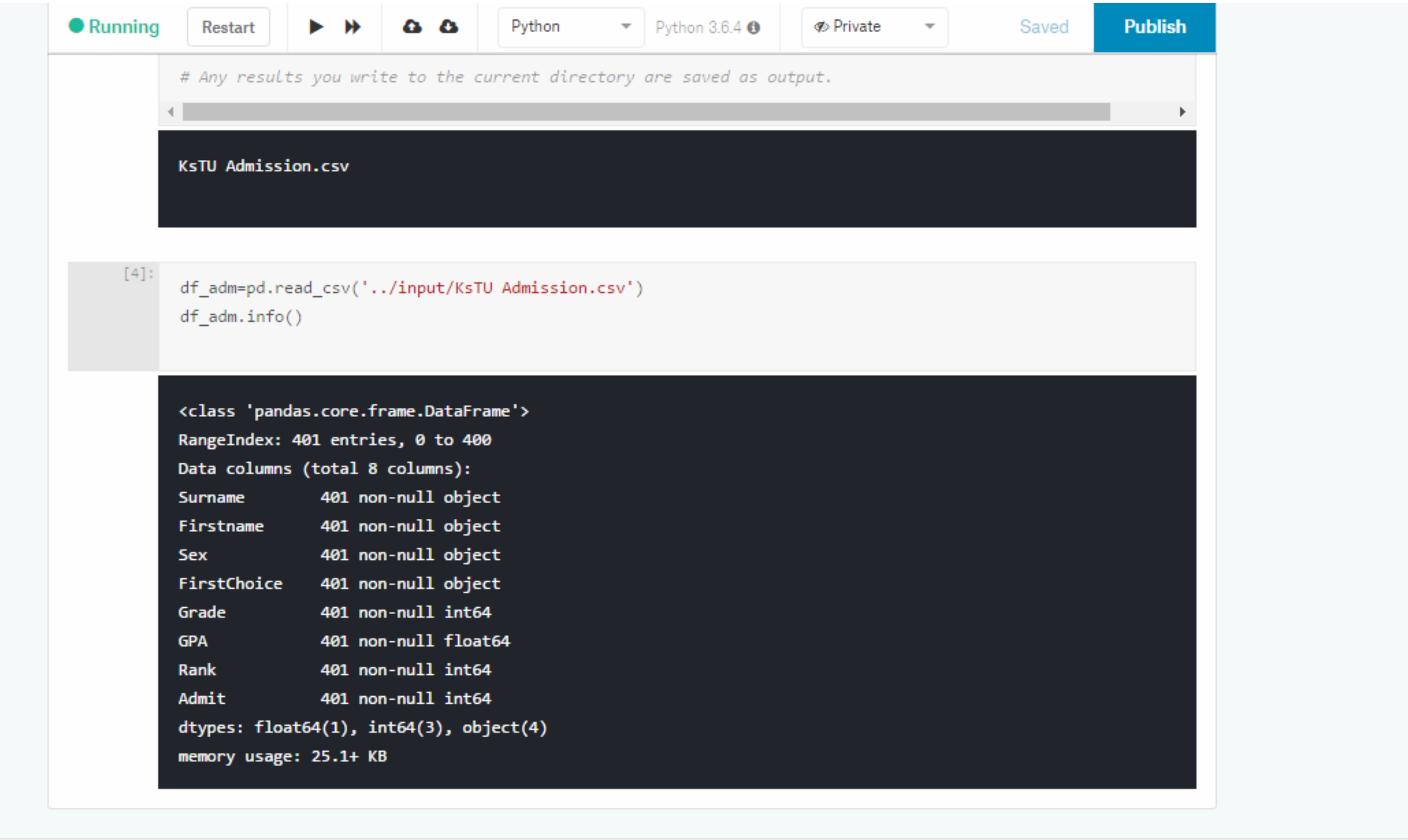
Figure 3 Admission dataset for statistics (see online version for colours)



```
[6]:  df_adm.head()
      df_adm.describe()
```

	Grade	GPA	Rank	Admit
count	401.000000	401.000000	401.000000	401.000000
mean	587.640898	3.389152	2.483791	0.316708
std	115.378121	0.380386	0.943590	0.465774
min	220.000000	2.260000	1.000000	0.000000
25%	520.000000	3.130000	2.000000	0.000000
50%	580.000000	3.390000	2.000000	0.000000
75%	660.000000	3.670000	3.000000	1.000000
max	800.000000	4.000000	4.000000	1.000000

Figure 4 The .info() command on admission dataset (see online version for colours)



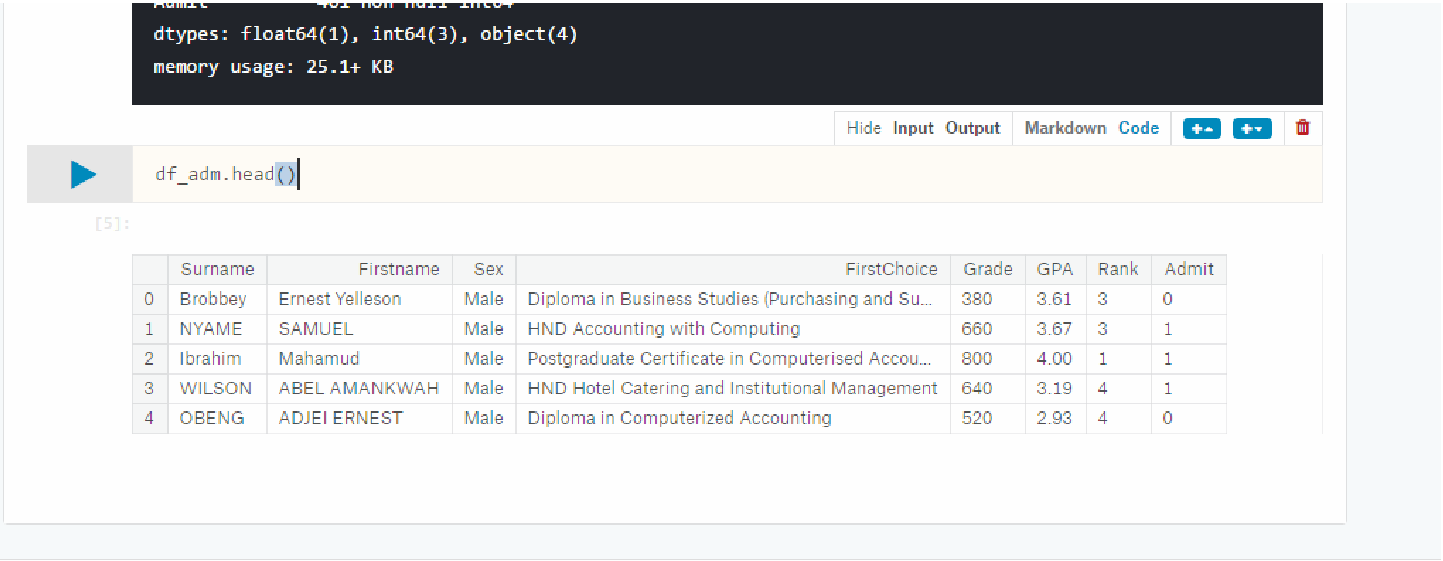
```
# Any results you write to the current directory are saved as output.
```

```
KsTU Admission.csv
```

```
[4]: df_adm=pd.read_csv('../input/KsTU Admission.csv')
     df_adm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 401 entries, 0 to 400
Data columns (total 8 columns):
Surname        401 non-null object
Firstname      401 non-null object
Sex            401 non-null object
FirstChoice    401 non-null object
Grade          401 non-null int64
GPA            401 non-null float64
Rank           401 non-null int64
Admit          401 non-null int64
dtypes: float64(1), int64(3), object(4)
memory usage: 25.1+ KB
```

Figure 5 Admission dataset information (see online version for colours)



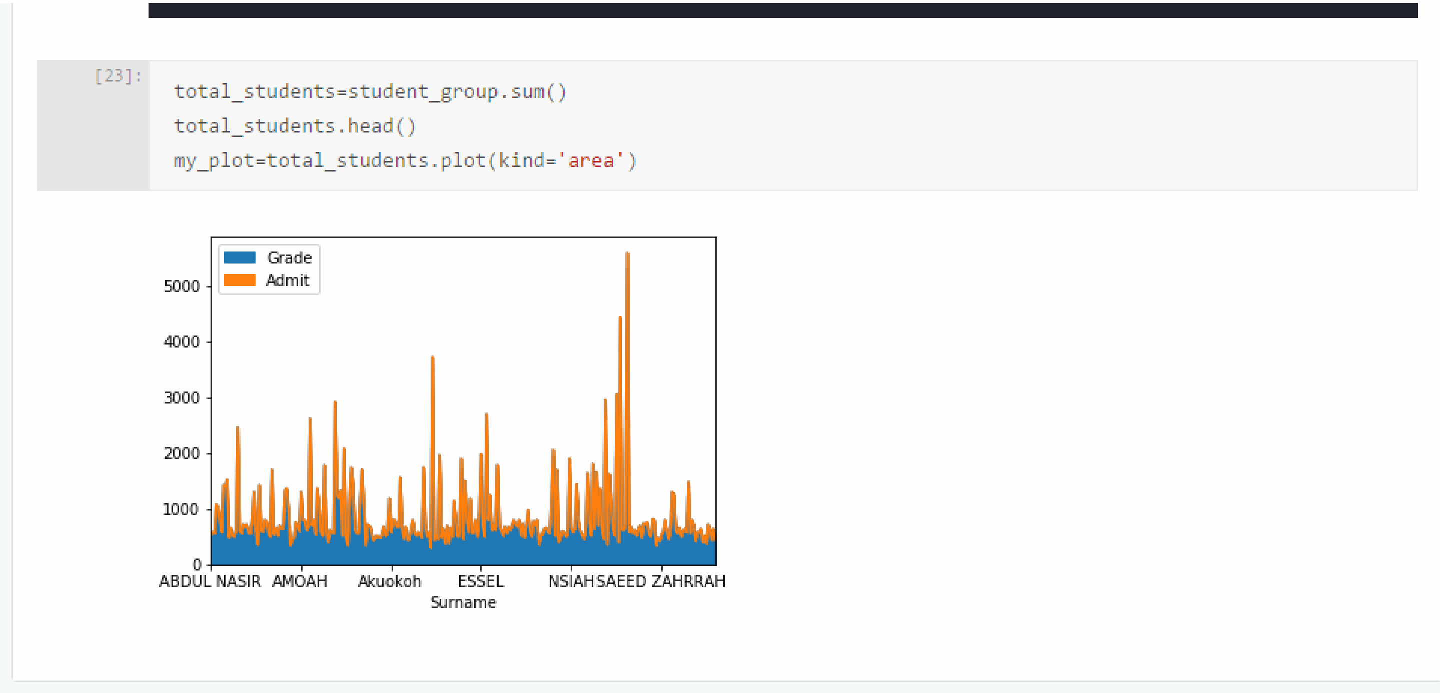
```
Admit            401 non-null int64
dtypes: float64(1), int64(3), object(4)
memory usage: 25.1+ KB
```

```
df_adm.head()
```

[5]:

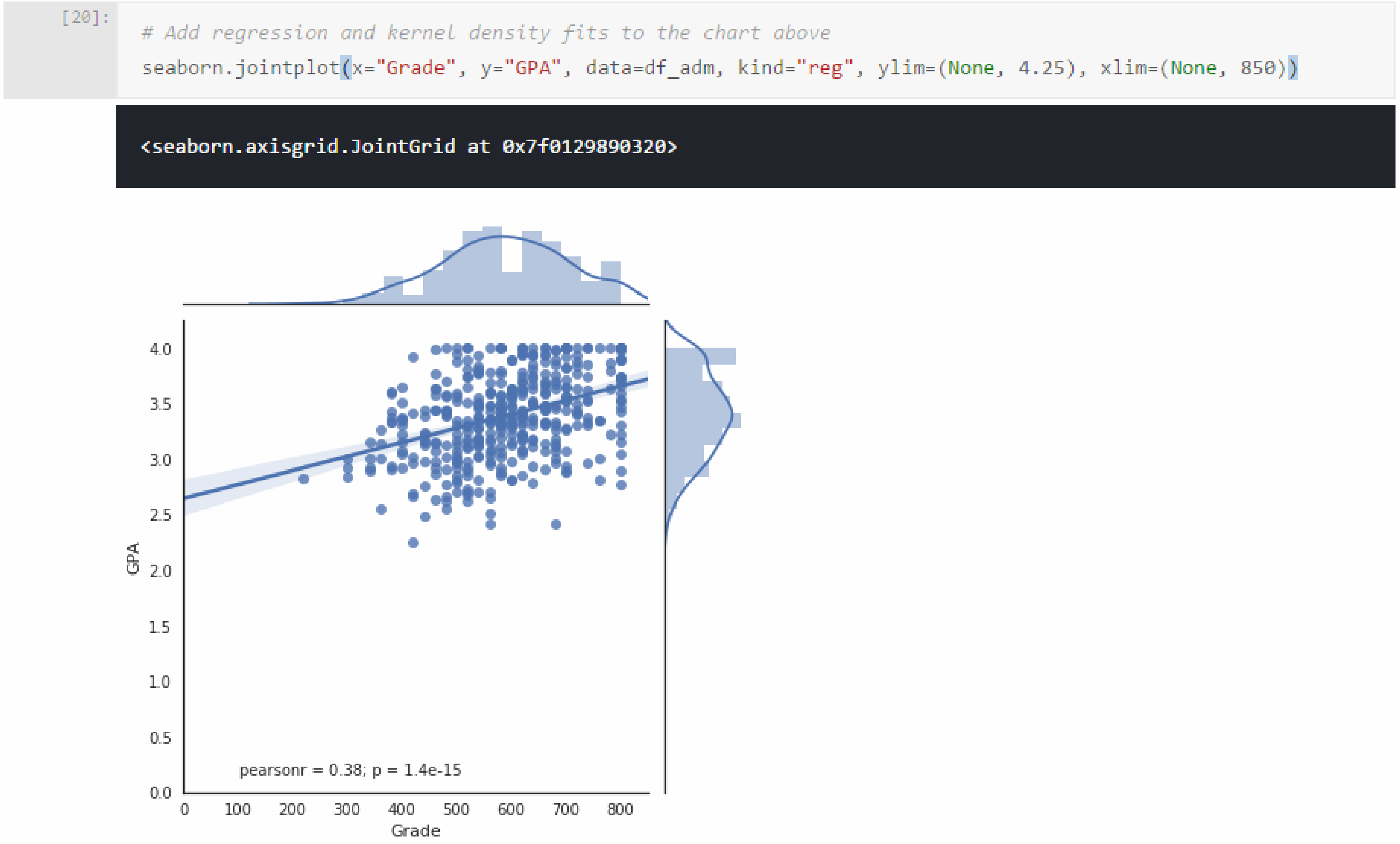
	Surname	Firstname	Sex	FirstChoice	Grade	GPA	Rank	Admit
0	Brobbey	Ernest Yelleson	Male	Diploma in Business Studies (Purchasing and Su...	380	3.61	3	0
1	NYAME	SAMUEL	Male	HND Accounting with Computing	660	3.67	3	1
2	Ibrahim	Mahamud	Male	Postgraduate Certificate in Computerised Accou...	800	4.00	1	1
3	WILSON	ABEL AMANKWAH	Male	HND Hotel Catering and Institutional Management	640	3.19	4	1
4	OBENG	ADJEI ERNEST	Male	Diploma in Computerized Accounting	520	2.93	4	0

Figure 6 The tendency of admission corresponding to grades (see online version for colours)



```
[23]: total_students=student_group.sum()
      total_students.head()
      my_plot=total_students.plot(kind='area')
```

Figure 7 Linear regression experiment on admission dataset (see online version for colours)



```
[20]: # Add regression and kernel density fits to the chart above
      seaborn.jointplot(x="Grade", y="GPA", data=df_adm, kind="reg", ylim=(None, 4.25), xlim=(None, 850))
```

```
<seaborn.axisgrid.JointGrid at 0x7f0129890320>
```

5 Experiment results

5.1 *Admission dataset processing*

The popularly known Apache Spark cloud platform, databricks was used to analyse and manipulate admission data from Kumasi Technical University. Figure 2 shows the connection of .csv file to the platform.

The statistics of the admission dataset is shown, which include the means, variance, max and min, etc.

The Apache Spark panda, code .info(), .head(), and .tail() was used to display the information on admission dataset. Figures 4–6 shows the .info() command.

Students are admitted to Kumasi Technical University based on their grade performance. The graph (Figure 6: the tendency of admission corresponding to grades) shows the tendency between the grades corresponding to admission.

6 Linear regression

Linear least squares are another simple linear method implemented in spark. Despite it was designed for regression, its output can be adapted for binary classification problems. Linear least squares follows the same minimisation formula described for SVMs and the same optimisation method (based on SGD), however, it uses squared loss (described below) and no regularisation method: $l_i = {}_{12}(\mathbf{w}T\mathbf{x}_i - y_i)_2$. The spark technology machine learning library was used to draw linear regression equation. The graph shows the regression prediction visualisation.

From Figure 7, the regression line, it indicates that students with GPA above 3.5 were admitted more. It also shows the admission depends on higher grade and higher GPA of candidates.

7 Experiment evaluation

The few commands used from pandas in python, which display Figures 4–7 show meaningful interpretation of information on data. Spark streaming and machine learning algorithm provide comprehensive and logical graphic representation of information/data than using Excel, Tableau and other statistical tools.

8 Conclusions

We demonstrate that data processing is an effective tool for educational sector, which can improve admission selection procedures and decisions. Most research papers focus on computational and theoretical aspect of education though little effort have been put on technological aspect of applying data mining techniques on admission selection. We therefore outline a simple spark streaming framework and machine learning algorithm to guide admission processing and focus on the number of students that can be admitted and rejected to reduce time and cost. The experiment conducted using spark streaming and machine learning algorithm confirm the proposed tools as cost effective and innovative tool for university students' admission data. The experiment show the practical usefulness of spark streaming and machine learning algorithm for data processing as it reduces time and cost and provides comprehensive graphical interpretation of data as compared to using Excel, Tableau and other statistical tools.

In future, we will be looking at using the proposed spark framework and machine learning to process students' attendance data every semester to access in the students' punctuality.

Acknowledgements

The work was supported by the National Natural Science Foundation (NSF) under grant (No. 61472294, No. 61672397), Key Laboratory of Spatial Data Mining and Information Sharing of Ministry of Education, Fuzhou university (No. 2016LSDMISO5) program for the High-end Talents of Hubei Province Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the above agencies.

References

'Spark Release 2.0.0' (2014) MLlib in R: SparkR MLlib APIs [..] PySpark MLlib algorithms, Python.

Baker, S. and Inventado, P.S. (2016) 'Educational data mining and learning analytics: potentials and possibilities for online education', in Veletsianos, G. (Ed.): *Emergence and Innovation in Digital Learning*, pp.83–98, DOI: 10.15215/aupress/9781771991490.01.

Brower, R. (2017) *Navigating the CBE Frontier: Leveraging Data to Improve Student Service and Outcome* [online] https://evolllution.com (accessed 3 March 2017).

Clark, R. (2017) *How Big Data Can Transform the Institution: Overcoming Obstacles and Achieving Change* [online] https://evolllution.com (accessed 3 March 2017).

Daniel, B.K. and Butson, R. (2013) 'Technology enhanced analytics (TEA) in higher education', *Proceedings of the International Conference on Educational Technologies*, Kuala Lumpur, Malaysia, 29 November–1 December, pp.89–96.

Hrabowski III, F.A., Suess, J. and Fritz, J. (2011) 'Analytics in institutional transformation', *EDUCAUSE Review* [online] https://net.educause.edu/ir/library/pdf/ERM1150.pdf (accessed 25 September 2018).

Jihong, X. and Yue, Z. (2017) 'Application and effects of blended learning in public courses in the university', *International Journal of Continuing Engineering Education and Life-Long Learning*, Vol. 27, Nos. 1/2, pp.87–100, DOI: 10.1504/IJCEELL.2017.081002.

Jomafuvwe, A.J. (2018) 'University access and 'comparative disadvantage' in Nigeria: a reflection on the criticality of equity for sustainable development', *International Journal of Higher Education and Sustainability*, Vol. 2, No. 1, pp.64–791, ISSN: 2056-4023; eISSN: 2056-4031.

Lavvitt, M. (2017) *Analytics and the New Academy* [online] https://evolllution.com (accessed 3 March 2017).

Long, P. and Siemen, G. (2014) 'Penetrating the fog: analytics in learning and education', *EDUCAUSE Review*, Vol. 46, No. 5, pp.30–40.

OECD (2013) *OECD Report: The State of Higher Education 2013* http://www.oecd.org/edu/imhe/thestateofhighereducation2013.htm (accessed 24 March 2014).

Okur, C. and Buyukkececi, M. (2014) 'Big data challenges in information engineering curriculum', *IEEE Network*, pp.5–12.

Raymond, J. (2014) 'Creating a learning organisation in higher education', *Industrial and Commercial Training*, Vol. 30, No. 1, pp.16–19.

Siemens, G. (2011) *How Data and Analytics Can Improve Education*, O'Reilly Media [online] https://www.oreilly.com/ideas/education-data-analytics-learning (accessed 30 October 2011).

White, C. (2011) 'Using big data for smarter decision making', *BI Research*.

Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Michael, J., Shenker, S. and Stoica, I. (2014a) 'Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing (PDF)', *USENIX Symp. Networked Systems Design and Implementation*.

Zaharia, M., Chowdhury, M.F., Michael, J., Shenker, S. and Stoica, I. (2014b) 'Spark: cluster computing with working sets (PDF)', *USENIX Workshop on Hot Topics in Cloud Computing (HotCloud)*.