Quality Assurance for Data Analytics

Rakesh Kumar, Birth Subhash, Maria Fatima, Waqas Mahmood
Faculty of Computer Science
Institute of Business Administration
Karachi, Pakistan

***Abstract*—Quality Assurance is a technique for ensuring the overall software quality suggested by Global Standards bodies like IEEE. The Quality Assurance for Data Analytics requires more time and a very different set of skills because Software Products, which are used for Data Analytics, are different than that of traditional ones. In result, these Software Products require more complex algorithms to operate and then for ensuring their quality, one needs more advanced techniques for handling these Software Products. According to our survey, Data Analytical Software Products require more work because of their more complex nature. One of the possible reasons can be the volume and variety of Data. On the same hand, this research emphasizes on testing of Data Analytical Software Products which have many issues because testing of these Software Products requires real data. However, every time the testing of these Software Products is based either on dummy data or simulations and these Software Products fail when they work in real time. For making these Software Products work well before and after deployment, we have to define certain Quality standards. In this way, we can get better result producing analytics Software Products for better results.**

Keywords*—Software Quality Assurance (SQA); data analytical softwares; data driven softwares; real time analytics; data analytics; quality issues; quality control

I. Introduction

Quality Assurance (QA) related activities are proven to be beneficial because it gives a confidence about the completion of requirements and needs that a user expect from a system, it can be the quality of product or it may be the accessibility and reliability, accessibility and reliability are just two qualities of systems, there are many more suggested by different experts. With change in time and access of internet, everything has changed its nature especially the field of information technology with increased usage of data driven softwares has impacted a lot. The more softwares will rely on data the more we will be facing challenges of software quality [1], [2].

In this paper we will not only discuss the quality issues in data driven softwares but we will also discuss best practices for testing of data analytics. Data analytics has contributed in almost every sector, especially for making life easier and for the betterment of our society, whether it is health sector or smart homes, data analytics has a huge impact on human race [3].

The increase in importance of data analytics has increased the work for developers especially for quality assurance teams because as we know and we have also practiced that Quality Assurance activities are the major portion of the effort in development of a software product. We have seen tremendous change in data related work and as a result the complexities and quality issues in software products have grown since few decades, these gaps in Software development must be handled with the help of performing activities to perform software quality assurance. Hence there is a need to define certain measures and approaches that must be followed to tackle quality issues of data driven softwares or data analytical softwares [4], [5].

Proliferation in the volume and variety of data has challenged practitioners to work more on the authenticity of data [6]. Let's discuss the internal structure and creation of data driven or data Analytical softwares because before talking about Quality Assurance we must know about the structure, development and purpose of those systems. Main goals of Data Analytical products are to read the data, understand the data and then find trends from it and most importantly to predict about the subject matter. These predictions can be pointing to a grater meaning of subject that is being observed, and results obtaining from these analysis can be leading to new theories and innovations, that is why these analytical softwares are getting familiarity day by day.

These softwares generally work in three stages: Data Preparation, Data understanding, and Prediction. This whole process continues after data gathering; data gathering can be done by doing surveys, focus groups or by taking interviews, etc.

- Data preparation that includes ETL (Extract, Transform and load), ETL includes data cleaning activities like, normalization, formatting the data into specified format and removing redundancies etc. Many times we need to make categories of continuous data.
- Data Understanding can be referred as Understanding of data by software, First software understands the data and then it is in the state to predict something by applying various predictive models.
- Prediction is the major part of any Analytical Software; it can be done after understanding of data by software and then applying some analytical method on it.

Quality Assurance is not only required for software development but it has many other benefits as well, software quality assurance also includes the contractual condition that are very essential when someone (an individual or an organization) is outsourcing a software; Q.A suggests best practices, rules and many other things like budget and deadlines to decide before signing the agreement [7].

This research focuses on the complexities in data analytical softwares and suggestions for quality enhancement in softwares [8].

II. Background

Availability of internet at every place especially in industries has a huge impact on the production of data, earlier data used to be stored in disks or hard drives now trend of storing data has moved towards digitization because data is producing in streams every second. Especially in the field of health where data is everything and to handle and manage this huge amount of data we need some special software that is totally different from traditional softwares. To develop tools/softwares that are data driven or we can say that are used for analysis purpose, we must define certain quality attributes because as we have already mentioned that these analytical softwares are very different and complex from traditional softwares [9].

Researchers from past have practiced many approaches to gather data and relevant information about the Research topic. Many methodologies and strategies to conduct a survey have also been proposed by practitioners and they have stressed the adaptation of various evidences related to research area [10]. With technological advancement and usage of social networking, the Data generating apps have introduced us with the big data and these apps generate such a huge amount of data that can be estimated as Terabytes, in a single day worldwide. Medical field also needs a quality assurance mechanism for their figure data as most of the clinical data contains medical figures, such as ex-rays. On an average a medium size hospital generates about 1 million figures per year [11]. To get some results from that data and its storage is not an easy task that’s why we need some special techniques and standards that can tackle this issue [12].

This research will be focusing on the data analytical softwares’ quality measures as we have seen changes in trend of softwares. This domain is still new and needs certain rules and standards so that quality of analytical software can be measured. We will be having a research survey along with a literature review, because research in this domain is not yet properly done so we will have to conduct a research survey too. The research survey will consists of questionnaire related to complexities and practices used in industries , this survey will be conducted from industry persons and students who have at least a bachelor degree in computer science because they are the one who will be working for those analytical softwares. Industry persons will also be taken interviews so that we can have a better knowledge about our research, since this is still a developing domain we will try to take surveys and interviews from people living outside of Pakistan.

III. Literature Review

Software Quality Assurance is a methodology that suggests globally accepted practices and standards to assure software quality. It aims to provide a quality product by conducting tests at different stages of Software project development [13].

Quality Assurance for data analysis needs domain knowledge in the distinctions of not only what can be the complexities, errors and efforts can be required at the time of collection of data, but interpretation of data can also be a huge challenge, because results from data can be misleading many times. That is why it is highly recommended that Quality Assurance team must be involved in the development phase and they should get in touch directly to the developers. These approaches can be followed when working on a project related to data Analytics:

Make a Quality Assurance group, who will verify that the output data that Data Analytical system produces is valid and give outcomes or results that we are supposed to get. Working with experienced individuals is necessary when we are working on some analytical system, because results can be dangerous many times. Adding Quality Assurance persons in developing team can be beneficial because they will be finding and resolving quality related issues at earliest and this approach can save out time, as a result, money can be saved. It is a mandatory task to plan for quality assurance activities, Quality Assurance team must make plans for testing; test plans can be interpreted as test cases etc. This approach gives a view to Quality assurance team and with this approach we can see a positive movement in our project testing side [4].

Testing in real environment is complex and money wasting technique, that’s why to simulate devices is a better option and this technique is successful so for, but in case of data analytics we need real data, like in case of traffic analysis we cannot simulate it or visualize it, though we can but it is not a successful technique. Recently Automated vehicle crash has occurred and this minor accident has caused company so much loss. In result, for data analytics we need real and variety of data so that beneficial output can be generated [14].

These approaches are very necessary because in majority of the systems data sources are too many and data is too huge to check it manually. For a simple query for example how many bike crossed this signal at 7:00 pm yesterday? Now this seems very simple in first sight but when we will be running this query manually it will take many hours may be a day or two in traditional or manual systems, that’s why we need to migrate to especial systems and we need to concentrate on quality more [15].

Many researches have highlighted the software quality assurance as a recent and autonomous field and it is also mentioned that software quality assurance is introduced after hardware quality assurance and similarities between these two domains are also discussed. Traditional methods are not enough to do the necessary improvements; there need to do some changes [16].

The purpose of this project is to highlight such quality attributes and techniques that must be present and used when developing or testing analytical software.

IV. Methodology

This research paper is based on literature review and questionnaire survey; we have conducted a research survey from variety of practitioners, those are either industry experts or students who have fresh and innovative ideas related to this domain. Motivation behind this survey was the unavailability of enough literature that could support our research [17].

V. Survey Responses

We have conducted a survey on Quality assurance for Data analytics. Through in this survey we have found so many different views. Some people think quality assurance for data analytics is really important while other thinks it is not important. Many people who have taken part in the survey are either employees or final year students. This summary will be showing the results from our survey.

We wanted to make sure that our audience must have CS background or at least they have enough knowledge about quality assurance so we asked them about their professions. People, who have given their views on this survey, are mostly students and developers. The graph summary in Fig. 1 showed that almost 60% are students, 30% are developers and remaining 10% are Instructors, Business Analyst and SAP consultants. Most students are graduates who have experience in this particular field. During this survey, we have analyzed the importance of quality assurance in every profession.

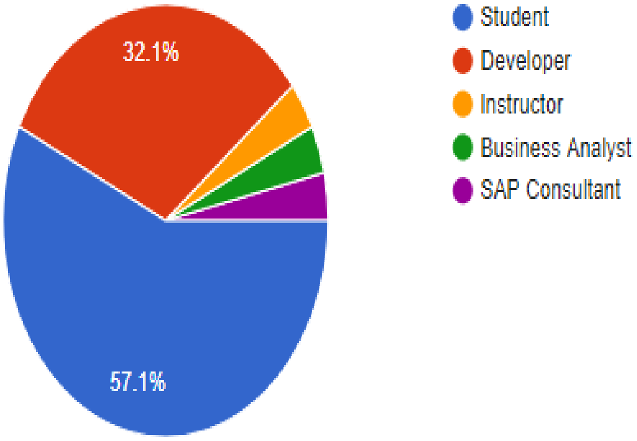

Fig. 1. Profession.

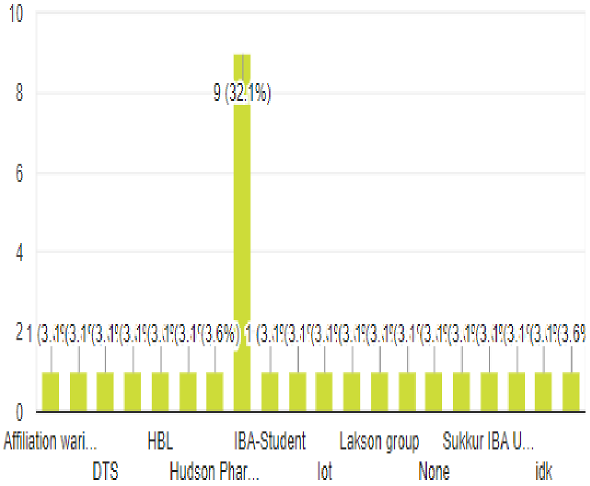

Fig. 2. Affiliations.

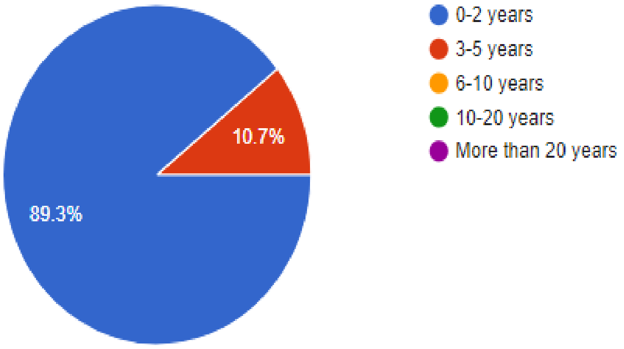

Fig. 3. Experience in years.

We also wanted to observe the people's affiliation; in which organization they are working or in which organizations they have previously worked. Fig. 2 shows that most people are either IBA students or they work in Hudson pharma Pvt. People from cloudKibo, Lot, HBL, Lakson group and Sukkur IBA have also helped us to analyze the need of quality assurance for data analytics. We can observe in this figure, the affiliation is on horizontal line and no: of employees on vertical line.

For this survey, it was very important for us to know about the work experience of the people who are taking part in this research, because mostly experienced people support quality assurance than fresh practitioners. In Fig. 3, we can see that most people have either no experience or less than three year experience. Graph summary shows that, only 12% are those who have 3 to 5 year experience and all remaining have 0 to 2 year experience. We tried to reach out to those employees who have more than six year experience but unfortunately we were only able to get 2 or 3 responses.

We wanted to know if people think that quality assurance is important for any software product. As it has already mentioned that most people who have taken part in this survey, are either final year students or employees because they have better idea of quality assurance rather than students who are in their first, second or third year. In Fig. 4 we can see, 96.2% of them think that it's really important and only 3.8% people think quality assurance is not much important. The reason behind this is maybe they don't have too much experience or maybe they are over confident on their code.

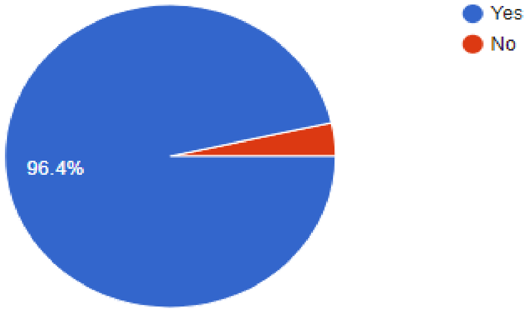

Fig. 4. According to you, is quality assurance important for any software product?

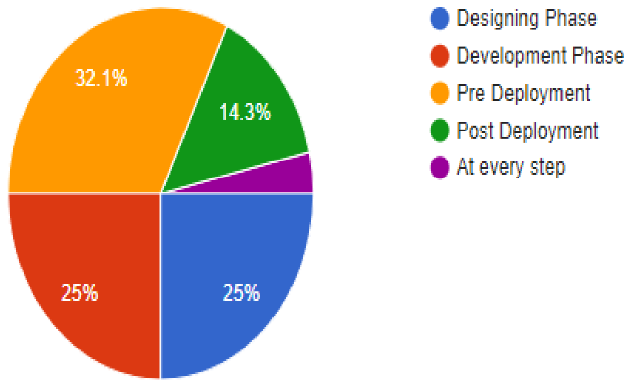

Fig. 5. At which stage of software development, quality assurance team should interfere?

We wanted to analyze, at which point quality assurance team should play its part. From Fig. 5, we can observe that everyone has its own view. 30.8% of them suggested that QA team should play its part before the deployment only, 15.4% of them suggested that they should apply tests and algorithms after deployment, 26.9% of them suggested that QA team should be there for only development phase and very few of them suggested that QA team should take part at every step and also after the deployment as well.

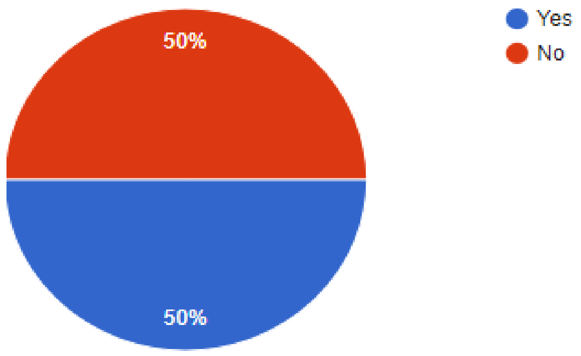

Fig. 6. Have you ever worked in the development of data analytical software.

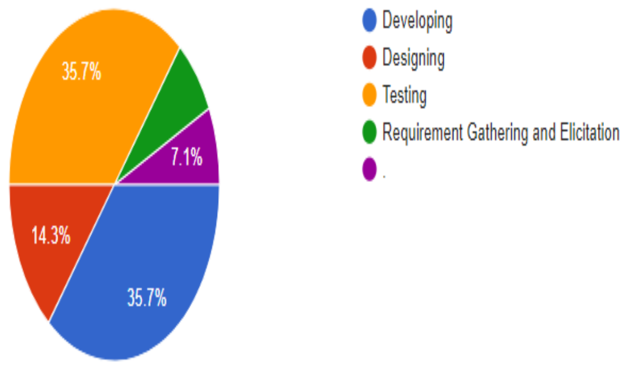

Fig. 7. If yes, what was your role?

We are conducting survey for data analytical software products that is why we wanted to observe our users in detail. Fig. 6, shows the responses of a question in which audience were asked, weather they have ever worked on any data analytical software. The summary told us that 50% of them have previously worked on data analytical software and they also believe that Quality assurance for Data analytics is really important.

From Fig. 7, we can understand that results are very interesting because question is answered by only those individuals who have ever worked for data analytical software and it is very important to know the views from them who have ever worked for an analytical software, almost 1/3 of the population have worked in the development of analytical software products and almost same number of Practitioners have worked in testing of those softwares, around 1/7 of the sample has worked in the designing phase of that software while a few people have worked in the requirement gathering phase of any analytical software ; requirement gathering can be included in designing phase.

From these results we can claim that almost half of the topic surveyors have worked in the creation of data driven or data analytical software products.

Fig. 8 gives us an overview about the opinion of people regarding the complexities of data driven or data analytical softwares.

The purpose of this question was to understand and find what people think about the complexities related to data analytical softwares. They were asked if complexities are different than the traditional softwares. Majority of the people that is around 4/5th of the total sample has accepted that complexities in data analytical system are more than the traditional softwares.

These results show that most of the Practitioners have clear understanding of data analytical softwares or at least they have a knowhow of these softwares. Those people who said the complexities aren't different it may be because they do not have ever worked for analytical softwares or maybe they have found it easy for them.

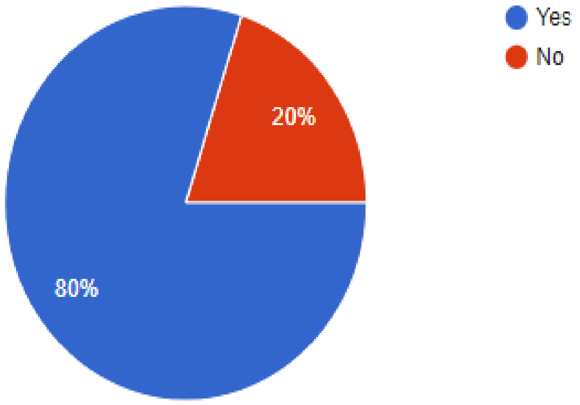

Fig. 8. In your opinion, are the complexities in data analytical softwares are different than the traditional softwares?

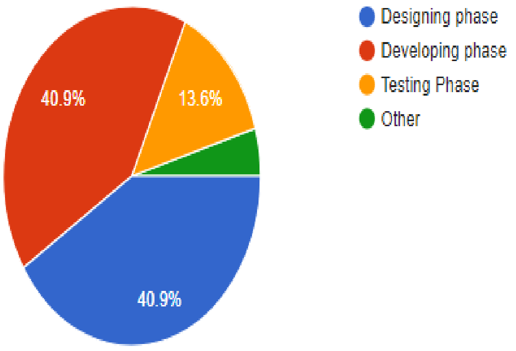

Fig. 9. If yes, in which phase these complexities occur?

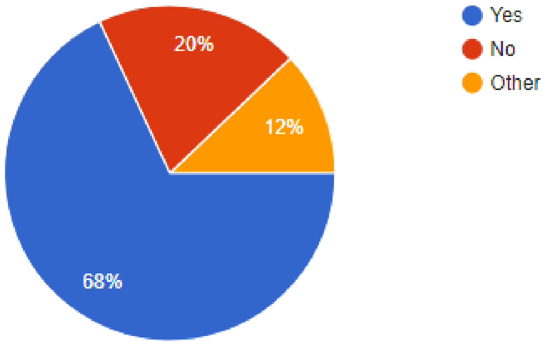

Fig. 10. Do you think there are any gaps in quality assurance of data analytical softwares?

This question was asked just to understand that during which phase practitioner find more complexities when working on analytical softwares. Fig. 9 explains, around 2/5th of the sample have said that they have found complexities during designing phase and I think this is genuine response because designing phase is the initial phase where software development team should be ready for upcoming complexities and should plan accordingly. Moreover almost same number of sample have expressed that they felt complexities during development phase that is also a genuine thing because if we do not plan and get ourselves ready during designing phase, we will having issues during development phase and almost 1/7th of sample feel that they face some complex situation during testing phase that is also understandable because testing of these softwares is also not an easy task, we need dummy data to simulate real life scenarios but many times real life results can be different than simulations , that means testing of these softwares is also a big issue. While few people also think that there can be complexities except these phases too.

Fig. 10 shows the responses of question “Do you think there are any gaps in quality Assurance of Data Analytical Softwares?” around 3/4th of the audience have said yes there are many gaps in the quality assurance of data analytical softwares, these gaps can be referred as barriers to test the analytical softwares, most important thing when testing these types of systems is data, because we have not sufficient data at the time of testing and in the end our software fails in real environment. While 1/5th of the population thinks that there isn’t any gap in the quality Assurance of these Analytical softwares it may be because they haven’t tried this task with their own hand or may be because they have find it easy or it is also a fact that they have find it interesting because many people find these analytical softwares very interesting. In addition to this 1/7th people have chosen others option it may be because they haven’t ever worked for these kind of softwares or maybe they have some different views about this.

Fig. 11 shows the summary of the responses, in which people were asked about the availability of the solutions related to these complexities, around 3/4th of the total population have positive opinion that there are many solutions to tackle these complexities. Although there are also few that was around 1/7th of population who do not know what to comment on this, that’s may be because they haven’t worked on any related software. Around 1/9th of population said no there is not any solution available that can tackle these complexities, that’s may be because they have worked on more complex softwares.

Fig. 12 shows the summary of the responses of a question, in which we asked our audience, if they allow us to contact for further details, around 3/5th of population is ready to tell us more about their experience related to this survey subject, few people said no, they do not want to be contacted and while 1/5th of the sample also chose other option that can be interpreted as they have some conditions before contacting them.

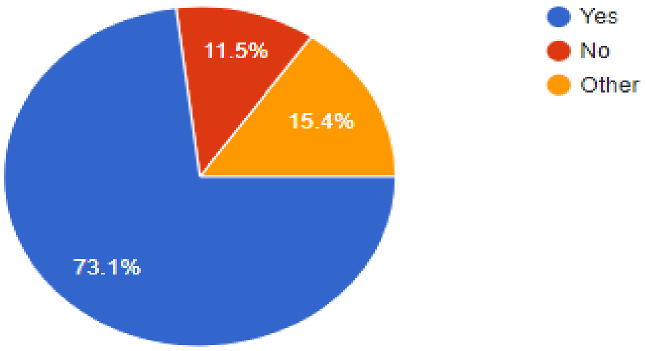

Fig. 11. Are there any solutions available to these complexities?

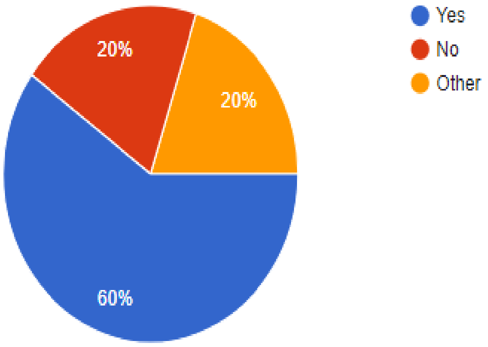

Fig. 12. For further information if we want to contact you personally, are you willing to help us in this regard?

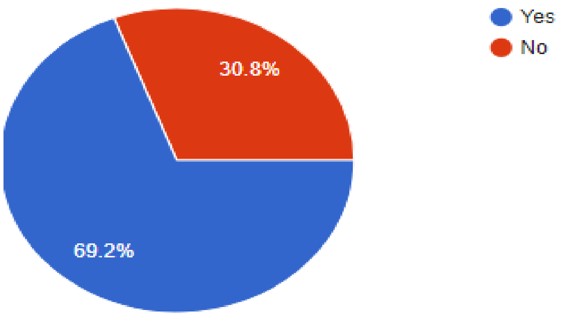

Fig. 13. Can we use your email address for contact purpose.

We also asked them if we can contact them on the same email address that they have provided in the beginning of survey form, Fig. 13 shows that most of them that was around 2/3rd of the population said Yes they can be followed on same email address, while 1/3rd of the sample doesn't want to be contacted on same email address that may be because they do not want to be contacted anymore.

This Survey gives us the idea what practitioners and concerned people think and has experienced while working on the data analytical softwares, all these responses will be helpful in our research and we appreciate each person who helped us achieve our goal.

VI. Conclusion

Quality Assurance for a software product is an essential part of development cycle and when it comes to Data Analytical Software, responsibility of Quality Assurance team increases due to the complexities because of such a huge amount of data and its variety. We have also conducted a survey to know about the views of practitioners those have worked/working in this domain. Majority of them have accepted that development of an Analytical software is more complex than a traditional software, and it was also answered that majority of them suffered issues during development of these software products while there are also few who suffered during testing phase. Our research concludes that there is a strict need of Quality Assurance team to interfere and suggest some solutions that can bridge that gap of quality requirements. Surveyors have also a positive feeling that these gaps can be filled by some new techniques and by applying those techniques, Quality Requirements for Data Analytical software can be fulfilled.

Reference

[1] M. Bruneforth, Martin and I. V. Mullis, Quality Assurance in Data Collection, vol. 10, M. O. Martin and I. V. Mullils, Eds., Boston: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College, 1996.

[2] L. Stahl, "Quality Assurance Project Plan for Data Analysis Activities for the National Study of Chemical Residues in Lake Fish Tissue," U.S. Environmental Protection Agency, Office of Science and Technology, Washington, D.C. 20460, 2007.

[3] F. Diko, Z. Alzoabi and m. Alnoukari, "Enhancing Education Quality Assurance Using Data Mining," 2016.

[4] "Quality Assurance for Analytics: 4 Steps to Avoid Big Headaches," 22 August 2017. [Online]. Available: https://www.qualitylogic.com/2017/08/22/quality-assurance-for-analytics/. [Accessed 04 April 2018].

[5] H. Foidl and M. Felderer, "Data Science Challenges to Improve Quality Assurance of Internet of Things Applications," in *International Symposium on Leveraging Applications of Formal Methods*, 2016.

[6] Narada Wickramage, "Quality assurance for data science: Making data science more scientific through engaging scientific method," in *Future Technologies Conference (FTC)*, San francisco,USA, 2016.

[7] D. Galin, Software Quality Assurance From theory to practice, Pearson education limited, 2004.

[8] J. Sargeant, "Qualitative Research Part II: Participants, Analysis, and Quality Assurance," *Journal of Graduate Medical Education,* vol. 4, pp. 1-3, March 2012.

[9] W. Raghupath and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Health Information Science and System,* p. 10, 2014.

[10] R. Per and M. Höst, "Guidelines for conducting and reporting case study research in software engineering," *Download PDF*, April 2009.

[11] V. S.Moustakis and L. Tsironis, "Knowledge Quality Assurance in Medical Data Mining," in *Proceedings of International Conference on Information Quality Management.*, Chania, Greece, 1996.

[12] C. Tao and J. Gao, "Quality Assurance for Big Data Application– Issues, Challenges, and Needs," *National Natural Science China,* p. 7, 2009.

[13] S. Farooqui and W. Mahmood, "A survey of Pakistan's SQA Paractices: a Comparative Study," in *29th International Business Information Management Association Conference*, Vienna, Austria, 2017.

[14] F. Lambert, "Tesla Autopilot confuses markings toward barrier in recreation of fatal Model X crash at exact same location," 3 April 2018. [Online]. Available: https://electrek.co/2018/04/03/tesla-autopilot-crash-barrier-markings-fatal-model-x-accident/. [Accessed 4 April 2018].

[15] SeattleDataGuy, "Data Quality Is Not as Sexy As Data Science," 15 September 2017. [Online]. Available: https://medium.com/@SeattleDataGuy/good-data-quality-is-key-for-great-data-science-and-analytics-ccfa18d0fff8. [Accessed 4 April 2018].

[16] F. J. Buckley and R. Poston, "Software Quality Assurance," *IEEE Transactions on Software Engineering,* pp. 36-41, 1984.

[17] R. M. Groves, F. J. Flower Jr., M. P. Couper, J. M. Lepkowski, E. Singer and R. Tourangeau, Survey methodology, 2 ed., John Wiley & Sons, 2009, p. 488.

Annexure-A

Questionnaire:

Email address:	. . . . .
Profession: Ex: Developer, Student	. . . . .
Affiliation: Ex: IBA, Arpatech etc	. . . . .

Question: Experience in Years?

- 0-2 years
- 3-5 years
- 6-10 years
- 10-20 years
- More than 20 years

Question: According to you, Is quality Assurance important for any software Product?

- yes
- No

Question: At which stage of software development, quality assurance team should interfere?

- Designing phase
- Development phase
- Pre-deployment
- post deployment
- Other. . . .

Question: Have you ever worked in the development of Data analytical Software?

- Yes
- No

Question: If yes, what was your role?

- Developing
- Designing
- Testing
- Others . . . . . .

Question: In your opinion, are the complexities In Data Analytical softwares are different than the traditional softwares?

- Yes
- No

Question: If yes, in which phase these complexities occur?

- Designing Phase
- Developing Phase
- Testing Phase
- Others . . . .

Question: Do you think there are gaps in Quality Assurance of Data Analytical Softwares?

- Yes
- No
- Others . . . . .

Question: Are there any solutions available to these complexities?

- Yes
- No
- Others. . . .

Question: For further information if we want to contact you personally, are you willing to help us in this regard?

- Yes
- No

Question: For further information if we want to contact you personally, are you willing to help us more in this regard?

- Yes
- No.

Question: Any suggestions or feedback?

Ans: