ISBN: 978-3-95450-193-9 doi:10.18429/JACoW-ICALEPCS2017-THPHA034

THE STUDY OF BIG DATA TOOLS USAGES IN SYNCHROTRONS

S. Alizada*, Imam Khomeini International University (IKIU), Qazvin, Iran
A. Khaleghi[1], Imam Khomeini International University (IKIU), Qazvin, Iran
[1]also at Iranian Light Source Facility (ILSF), Tehran, Iran

Abstract

In today's world, there is plenty of data being generated from various sources in different areas across economics, engineering and science. For instance, accelerators are able to generate 3 PB data just in one experiment. Synchrotrons industry is an example of the volume and velocity of data which data is too big to be analyzed at once. While some light sources can deal with 11 PB, they confront with data problems. The explosion of data become an important and serious issue in today's synchrotrons world. Totally, these data problems pose in different fields like storage, analytics, visualisation, monitoring and controlling. To override these problems, they prefer HDF5, grid computing, cloud computing and Hadoop/Hbase and NoSQL. Recently, bigdata takes a lot of attention from academic and industry places. We are looking for an appropriate and feasible solution for data issues in ILSF basically. Contemplating on Hadoop and other up-to-date tools and components is not out of mind as a stable solution.

INTRODUCTION

The world's most valuable resource is not oil anymore, it is "Data", the oil of digital age. Nowadays there is a plenty of data being generated from various sources in different areas such as economy, engineering, science, etc. 5 companies (Google, Amazon, Apple, Facebook and Microsoft) are the most well-known companies in the world and the reason is "data" [1]. How useful could be this amount of data? Have you ever thought about this? Ending world hunger, reducing crime, halting deadly outbreaks of diseases are a small advantage of using data to solve problems. For instance, $300 billion to $450 billion could be saved by using data for a better outbreak prediction just in the United States [2][3]. The five companies named above have earned totally $25 billion profit in the first quarter of 2017 [1]. Although data innovation itself isn't the solution in most cases, it could present an extremely fundamental piece of information to unlock the doors to the new answers. Todays, the sheer volume of available data is growing rapidly. This growing data is getting too much and we need to dwindle it down to a meaningful subset. A new technology has appeared and has aroused great expectations and that is "Big Data".

Big Data is a term used to describe the collecting, processing and making available huge volume of streaming data in real-time. The three V's are Volume, Velocity and Variety with credit to Doug Laney [4]. Having a lot of data which are pouring into your organization is one side of the coin, being able to store, analyse and visualize the mentioned data in real-time is a whole different thing, the other and the most important side of the same coin.

In this project, we are trying to have a complete study on big data tools and tested techniques in various light sources around the world for data in beamlines, emphasizing on the storage and analytics aspects.

SYNCHROTRONS DATA ISSUES

In many scientific fields such as physics, statistical analysis of large data sets is common. The particle accelerator's data offers a good example of scientific data. Particle accelerators generate data at a rate of 1 MB per collision event, and such events happen at a rate of about 600 million per second. Handling this huge amount of data is a core problem and the solution is Big Data [5]. For instance, accelerators are able to generate 3 Petabyte data just in one experiment [6]. Synchrotrons industry's data is an example of the volume and velocity of data. Synchrotron's data is too big to be analyzed at once. Though some light sources can deal with 11 Petabyte, they confront with data problems [6]. The explosion of data becomes an important and serious issue in today's synchrotrons world.

Totally, these data problems pose in different fields like storage, analytics, visualisation, control and monitoring. To cope with these problems, they prefer HDF5, grid computing, cloud computing and Hadoop/Hbase and NoSQL [7]. Recently, Big Data has attracted lots of attention from academic and industrial organizations. We are basically looking for an appropriate and feasible solution for data issues at ILSF. Contemplating Hadoop and other up-to-date tools and components is not out of mind as a long term solution.

OUR APPROACHES

In this regard, an interview was designed and sent to the interviewees after validation procedures. This interview has 10 sections. Which is shown in the Figure 1. Each of these sections is for an interview. This interview is structurally a semi-structured interview.

Given that we did not have direct access to the interviewers and it was not possible to set a specific time for online interviews on Skype, so I came across interviews and online forms which I encountered with many forms. Examining the advantages and disadvantages, in fact, assessing which online forms would provide more opportunities for our interview and could meet our demands, was our next step in this study.

* s.alizadeh@eng.ikiu.ac.ir

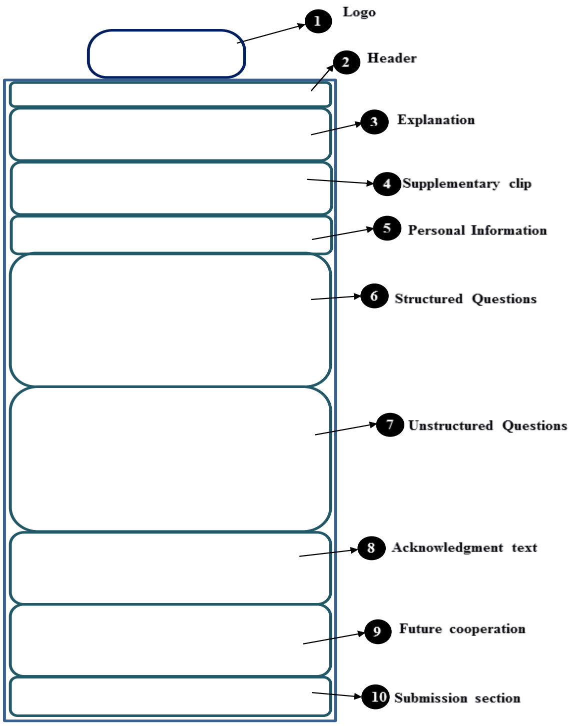


Figure 1: The interview structure and sections.

In some of the questions in this interview, we needed a list of experiments which the list was prepared by two groups during the periodic sessions was shown in the Figure 2. The two groups include beamline scientists and synchrotron hardware experts.

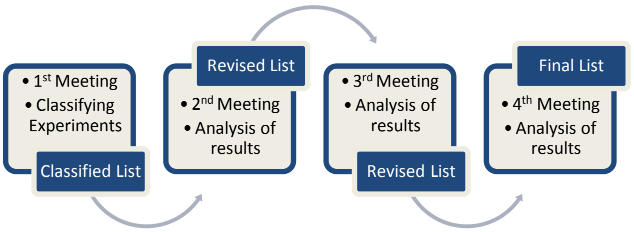


Figure 2: Periodic sessions for experiment list.

Online Interview Form Designed. Every one of the scientists and statistical population was selected and 583 beamline scientist and experts were selected for this research, as well as officials and managers of the control and computer department and senior managers of each beamline. An email was sent for participating in this research.

This statistical population has been selected from the following synchrotrons around the world. The list of synchrotrons has come in the Table 1.

Table 1: Participated Synchrotron List

Name	Country	Name	Country
ALBA	Spain	INDUS1/INDUS2	India
ANKA	Germany	LNLS	Brazil
AS	Australia	MAXIV	Sweden
CLS	Canada	NSLS	USA
DESY	Germany	PAL	Korea
DLS	UK	SAGA-LS	Japan
ELETTRA	Italy	SLRI	Thailand
ESRF	France	SLS	Switzerland

According to the information obtained, we arrive at the following conceptual model (Figure 3). Finding relationship between each beamlines and data management aspect was our main task in this research.

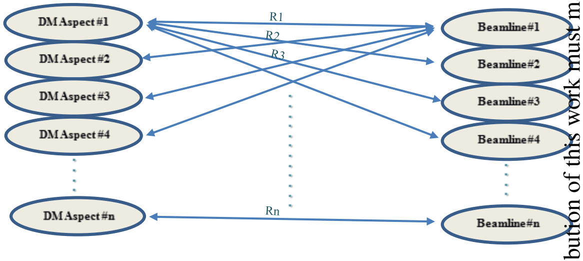


Figure 3: Conceptual model.

RESEARCH RESULTS

In this part you can see the results, we have obtained by the interviews which were held. Chart 1-2-3-4 and Figure 4 are representing the final results.

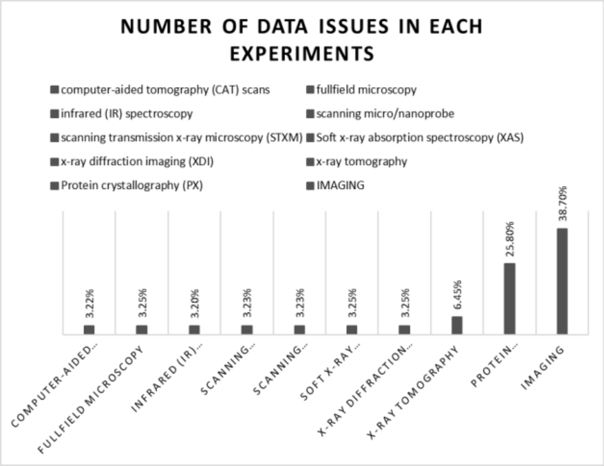
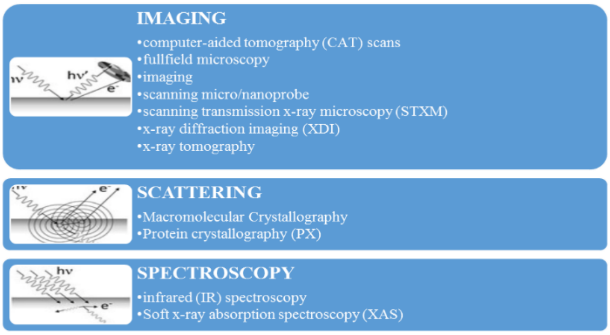


Chart 1: Importance of DM aspects based on experiments.



IMAGING
- computer-aided tomography (CAT) scans
- fullfield microscopy
- imaging
- scanning micro/nanoprobe
- scanning transmission x-ray microscopy (STXM)
- x-ray diffraction imaging (XDI)
- x-ray tomography

SCATTERING
- Macromolecular Crystallography
- Protein crystallography (PX)

SPECTROSCOPY
- infrared (IR) spectroscopy
- Soft x-ray absorption spectroscopy (XAS)

Figure 4: Experiments classification based on techniques.

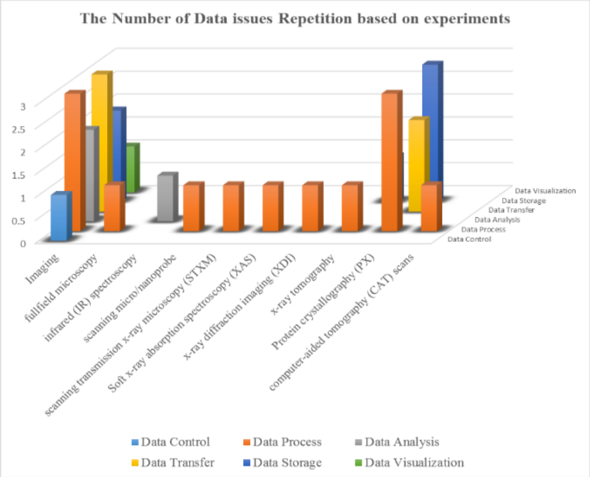


Chart 2: Number of data issues on experiments.

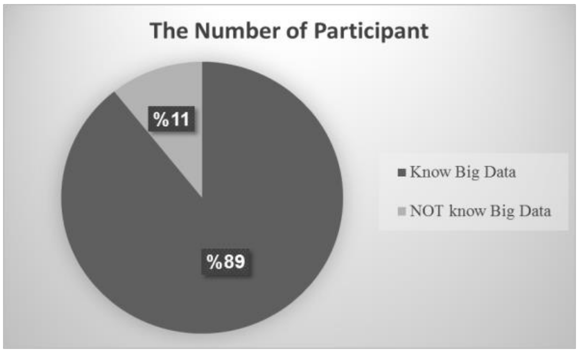


Chart 3: Amount of familiarity with Big Data.

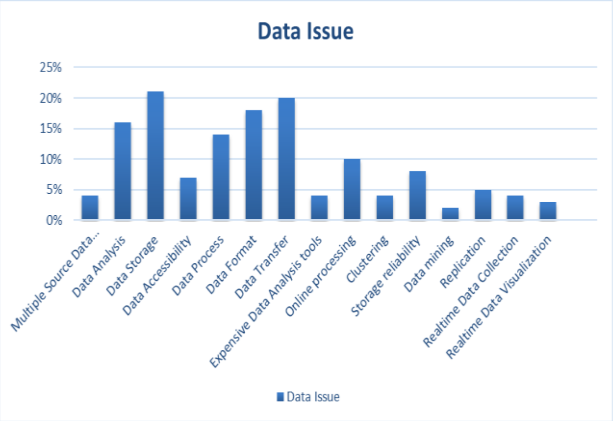


Chart 4: Raised issues by the interviewees.

CONCLUSIONS

Based on the conceptual model of the research, we came to the following model at Figure 5 after analyzing the results.

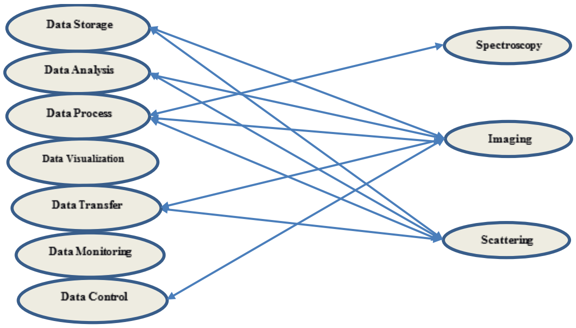


Figure 5: Final conceptual model.

SUGGESTED FUTURE WORK

Since the information obtained in this study is basic information and can be used in decision making in many future researches. I have several suggestions for continuing this research that I've listed below:

- Designing a big data framework that is specially for synchrotrons can now be for a specific section or for a specific beamline. This offer is based on the information provided. In many interviews, there is a mismatch between formats or software and even the lack of platforms. By integrating and coordinating these softwares in a particular framework, the problems of the data format of the scientists will be solved.

- In this dissertation, we outlined some of the obvious problems in some of the beamlines, which, in the follow-up to this research, can be used to personalize and identify some big data tools.

- This research has been done on the synchrotron industry. It can be applied to any other industry, including banking and economics, astronomy, and astronomy, to find the application of big data tools.

ACKNOWLEDGEMENT

We would like to express our deepest appreciation to all the members of Synchrotron Light Source Community who helped us in the process of this research by willingly sharing their precious time to answer the survey Questions. we are also grateful to prof. Rahighi, the manager of ILSF and control and computer divisions' staff in ILSF for sharing their pearls of wisdom with us during the research.

REFERENCES

[1] The World's Most Valuable Brands, `https://www.forbes.com/powerfulbrands/list`

[2] Tanza Loudenback, "The 10 most critical problems in the world, according to millennials", Aug. 23, 2016.

[3] How Data is Solving some of the World's Biggest Problems, `http://www.ozy.com/`

[4] Laney, Doug Management, Data Volume, Controlling Data, "Application Delivery Strategies", META Delta Group, 6 February 2001.

[5] Olof Barring, CERN, IT Department et al, CERN openlab Whitepaper on Future IT Challenges in Scientific Research, May 2014, CERN openlab.

[6] Diamond's Big Data, `http://www.diamond.ac.uk/Home/News/LatestFeatures/15_05__15.html`

[7] N. Kikuzawa, *et al.*, "Status of Operation Data Archiving System Using Hadoop/HBase for J-PARC", *Proceedings of PCaPAC2014*, Karlsruhe, Germany, 2014, paper FPO016, ISBN 978-3-95450-146-5193.