Insights into the software Industry - Trends and Preferences

Akash Sadanand Hande
School of Computing
National College of Ireland
X17156220@student.ncirl.ie

Abstract — **There are many different views on the preferences of developers towards tools and trends in the software industry. In 2017, major players in the industry such as Deloitte have published their views on technology. The basic purpose of such analysis is to know insights of the software industry and to improve the process of software development to meet business expectations. While thinking as a third party, this type of analysis is held in the organization and they have boundaries of company laws. After considering the software industry is not dependent on one organization and their reports such kind of analytics should be held on a global level without any boundary of the country, organization, or age group.**

This report describes the quantitative analysis of stack overflow annual survey to know technological insights in the software industry. Analytics in this report includes multiple objectives ranging from popular version control tool to employment status in the exciting aspects of Artificial Intelligence. This analysis has been carried out by using two different distributed frameworks known as Hive and Pig.

Keywords—Data Analytics, Stack overflow, Hadoop Ecosystem, Pig, Hive

I. Introduction

In today's digital generation of software application where each application from various domain such as commerce, social media, business or financial etc. producing a huge amount of data ranging from exabytes to zettabytes. When it comes to data analytics, where analyzing and processing of data is done for extracting valuable information such as classification and prediction, the traditional frameworks may get fail. For analytics of such huge amount of data, we require distributed frameworks such as Hadoop. Research nowadays focus on the implementation of various frameworks for data analytics which includes MapReduce, Hive, Pig etc. In this paper, different frameworks have been utilized to get knowledgeable information which can be used for prediction.

Each year, Stack Overflow ask developers community about everything from their favorite technologies to their job preferences. This dataset can be useful to plot some pivotal values to thinking and preferences of developers towards trending technologies. The process is to extract, clean, transform, load the data from the source database to the distributed database using the Hadoop framework. During the data analytics process, the key information is derived by using different Hive, Pig jobs. Valuable knowledge is visualized to derive business values using Tableau. These parameters can be useful to drive the developer's community for the technological growth.

A. *Objective of Analysis*

There are majorly six objectives of analyzing [1] dataset as above. 1. Know programming trends in Web Development. 2. Find out popular version control tools for Agile methodology. 3. Compare employment status in 2017 and 2018. 4. Check popularity of Stack Overflow amongst age groups in United States 5. Find out an exciting aspect of increasingly advanced AI from a developer's point of view 6. Job satisfaction of Full-stack developer.

Business Objective Question: Can we use Stack Overflow annual survey results to understand their popularity, trending technologies amongst developers around the globe and exciting aspect of increasing technologies such as Artificial Intelligence?

B. *The motivation of the Problem*

Just as "Big Data Analytics" has become more and more important to use in this age of information, we can determine various ways through which we can make good decisive steps to achieve positive results. All great minds from the computing community use massively generating data on daily basis to formulate better strategies for the future of their business, research, and work. As part of this project as well as programmer's community, I got an opportunity to work with real-time data set to find out some insights of the developer's community. It was challenging to work with real-time data set as it was generated through an annual survey held amongst developers around the globe and this analysis includes two recent surveys i.e. 2017 and 2018.

C. *The relevance of chosen topic*

The chosen topic is quotative analysis of Stack Overflow annual survey published in the year 2017 and 2018. This includes opinions over 100,000 developers around the world from their job preferences to their favorite technologies. There are many things we can learn from such a rich, large data set and we can explore about various topics such as developer's salaries, remote work, the impact of emerging technologies on developer's community, the popularity of Stack Overflow etc. Since this data set has millions of rows and numerous numbers of columns HADOOP is the best approach for analytics of this data set.

D. *Detailed Research Question*

The project is based on real-time use case formed for academic research purpose on the data set published by Stack Overflow. The use case formed is more focused towards finding insights of developer's community for technological trends micro focusing to tools and programming languages, comparison of employment status, the popularity of Stack Overflow amongst different age group and motivation of developers towards Artificial Intelligence from developer's point of view. These five objectives form the paper's five research questions:

- Which are the top Programming trends in Web Development?
- Which are the top version control tools for Agile methodology?
- Can we compare employment status in 2017 and 2018?
- What is the demographic popularity rate of Stack Overflow in the United States?
- What motivates developers' interest in advanced AI?
- Job satisfaction of full stack developer?

As published data from both sources [1][2] have uncleaned data rows and as formed use case expects valid data format. Based on the use case and research question, raw data has been filtered and then processed to meet the requirement of use cases.

II. RELATED WORK

Trends are very fluctuating in the Software Industry and software developers do not attract easily to a new tool, to conclude this several academic, as well as industrial research, has been conducted. Christof Ebert and Kris Shankar from Vector Consulting Services and Microsoft respectively represent trends, topics, and technologies relevant in 2017 and as part of leading software giants they show the heading of innovative products and solution development [3]. As part of their research, they conducted a survey of their worldwide clients and which includes about 1500 decision makers from different regions and industries. To keep survey simple and traceable they targeted mid-term and short-term trends in the software industry. Combination of innovation, efficiency and digital transformation is clear managerial outlook in software industry whereas security and safety are the pillars of mid-term trends. The research held in [4] by Mohammad S Raunak and David Binkley highlights "Agile and Other trends in Software Engineering" which include Tools usage in the Agile development process, trending software development process in the industry, success and failure indicators etc. For this analysis, they conducted a survey amongst 99 software developers and managers from different companies asking their choice of process, tools, and techniques. Their study suggests that agile adoption has been dominated by an adapted version of scrum and key of successful implementation of the agile process is a selection of better development tool.

Economic and social issues have been analyzed to point out software engineering challenges in the coming years in [5]. Since the software industry needs to treat issues beyond technical perspective they map on what is currently known about the software industry and plot three-dimensional (social, business and technical) viewpoint towards software industry. Their observation clearly shows that software ecosystems research is mainly focused towards ecosystem modeling, open source software, and business issues.

From the second point of view, Stack Overflow is a huge source of information for the industry as well as academics practitioners. There are multiple papers that have taken this 'platform' as an analyzing platform to get useful insights into practice adopted by software developers. In 2017 [6], Tanveer Ahmed and Abhishek Srivastava have used this platform to explore the habits of users. More specifically they have found out the behavior of technical users online and tried to present the human aspect of the online community. For this project, they mine stack overflow repositories to find few patterns of human behavior which highlight the real-world impact of having open call-based software engineering platforms. They presented their analysis in the graphical format.

III. METHODOLOGY

This study took public data set provided by Stack Overflow for academic research purpose, Data cleaning is done through basic feature provided by Microsoft Excel, processing is carried out by using Hive and Pig and at the end, results have been highlighted by using Tableau. Here is the detailed explanation of the methodology:

A. *Choice of Dataset*

Each year from 2011, Stack Overflow carries out the annual survey and opens the opportunity for data analysts to learn preferences of developers from world's most trusted and largest community of professional developers. They publish their survey results on their public site [2], as well as they, partnered with Kaggle which is a most popular platform for predictive modeling and analytics competitions for data miners for publication of dataset [1]. This dataset has been used to learn the preferences of software developer's community towards specific programming languages, tools, employment status and emerging technologies such as AI.

For this project two recent datasets i.e. 2017 and 2018 has been used for the analysis purpose. In 2018 survey dataset we have 98,855 responses for 129 different questions and for 2017 we have 51,392 responses for 154 questions. As both the datasets are published from the same genuine source with the help of a large community of software developers we can get opinions from all over the globe and which is very useful for this analysis.

B. *Data Processing Activities*

Data processing activities in this project include data cleaning, processing using different frameworks and generating results as explained below:

1) *Data Cleaning*

Data cleansing or cleaning is the first major step in the overall data preparation process which is a process of identifying, analyzing and correcting raw and messy data. Inadequate data cleaning and preparation leads major cracks in data analytics and analysis can be failed. The dataset used for this project is in csv form is much more structured but have some corrupted values which may distort analysis, so we standardized it into a single format. We enrich the data after removing the corrupted data like special characters for better analysis, validate it to improve data quality and inconsistency issues after all of this we used clean data for further processing.

2) Data Analytics

Data Analytics involves applying mechanical or algorithmic process to derive insights and draw a conclusion about the information. To achieve the objective of this project we use two major technologies i.e. Hive and Pig for the processing of such a huge data and generate output in .txt format. After that, for the better understanding and visualization purpose, we used "Tableau" which is data visualization software focused on business intelligence [9]. We generated multiple charts and graphs for the better understanding of outcome.

C. Choice of Technologies

In the Data Science community, list of technology vendors offering big data solutions is apparently infinite. Many of the big data platforms are popular right now such as the Hadoop Ecosystem, R, Spark, In-Memory databases etc. From this wide range of big data solutions, we choose Pig and Hive from Hadoop Ecosystem for the analysis purpose. Apache Pig is the platform for analyzing big data sets which consist of a high-level language for expressing data analysis programs. A most important feature of Pig programs is their simple structure to parallelization which is more helpful to work with large datasets [10].

Another tool used for analytics purpose in this project is Apache Hive which is data warehouse software facilitates managing large datasets with reading, writing operations in distributed storage using SQL [10]. Both the tools are part of Hadoop Ecosystem and are widely used for the Data Analytics. As mentioned earlier, for the visualization purpose we used Tableau.

D. Implementation of Algorithm

As mentioned in the objective of analytics section five objectives has been chosen to formulate the use case of this project. Implementation of those are explained below:

1) Programming trends in Web Development

The saying "the only constant changes" seems to be an industry-defining in Web Development. Web Development is changing every second from the evolution and 2018 will be no different. Analysis of this objective is carried out through the implementation of pig algorithm. objective1.pig contains a pig script which implements counting with counters, cartesian product, filtering pattern and top 10 design pattern. By using this script top 10 programming trends in web development have been summarized.

2) Popular version control tools for Agile methodology

The hottest question for software development teams who are working in Agile methodology is "Who done this?" and to answer such knocking question senior members in the team use version control. The key to success in agile software development process is the implementation of versioning controls such as SVN and Git. Analysis of this objective is done through the implementation of "Hive SQL Query" in objective2.hql. For this implementation "counting with counters" summarization design pattern has been used which utilizes MapReduce Framework's counter utility to know the sum of entirely on map side without generating any output [10]. Apart from that "Cartesian Product" and "Filtering Pattern" has been used to form grouped result and focusing results on developer type and methodology used by the software professional [10].

3) Comparison of employment status in 2017 and 2018

To know the employment rate along with employment type (Full-time, Part-time, Unemployment) we carried out this analysis. For the implementation of this objective, we use Pig Script along with Composite join which is a specialized type of join operation performed on the map side with multiple, very large formatted input datasets [10]. As we are comparing two data sets i.e. 2017 and 2018 we need full outer join and composite join is the best choice for this implementation.

4) The popularity of Stack Overflow amongst various age groups

The hidden power of Stack Overflow is its popularity amongst different age group across the globe. To know the popularity of Stack Overflow we carried out this analysis with the help of Pig algorithm in objective4.pig. The query implements three design patterns namely "Cartesian Product", "Filtering Pattern" and "Counting with Counters" of Summarization pattern [10]. The cartesian product is an effective way to pair with every other record from the input. Counting with counters is an efficient way to retrieve count summarization of large dataset. Filtering pattern is useful for keeping records of interest and removes the non-interested records. As this analysis is based on relationships between all pairs of individual records, objective requires a count of age group from the United States three design patterns have been implemented.

5) The motivation for developers towards advanced AI

It is somewhere safe to predict that advanced AI will continue to top of the trending technologies in 2018 but what is the basic reason for that? To know the base platform from a developer's point of view we are carrying out this analysis by using Hive SQL Query in objective5.hql. This query implements three different design patterns namely "Cartesian Product", "Counting with Counters" of Summarization pattern and "Filtering Pattern" [10]. Filtering pattern is used to filter out the records which we are not interested and keeps major four categories in which objective is rest on. As an objective for this analysis wants to analyze the relationship between all pairs of individual records as well as a count of similar record group cartesian product and Counting with counters has been applied.

6) Job satisfaction of Full stack developer

Full stack developer considered as good profile amongst developers' community as you can get no of technologies to work with from UI layer till DB layer of the application. To find out job satisfaction of full stack developer objective6.pig has been written. Which implements "Filtering pattern", "Cartesian Product", and "Counting with Counters" of Summarization pattern [10].

IV. RESULT

The source data contains about a million rows and to meet all mentioned objectives combination of Hive SQL Query and Pig Script has been developed. The result of these scripts we generated in .csv files and visualized by using Tableau [10] in form of different graphs and charts.

1) Programming trends in Web Development

As we can see in the results C#, JavaScript and Java hold top positions for the programming languages for the programmers in the world.

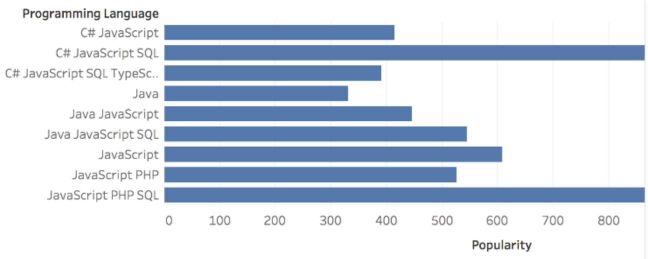

Figure 1: Programming Trends

2) Popular version control tools for Agile methodology

Unsurprisingly, GIT holds a topmost position in the popular version control tool in agile methodology. As Git is providing a faster release cycle, community development it is the key to the agile development process where multiple developers must work parallelly and deliver the project with the iteration.

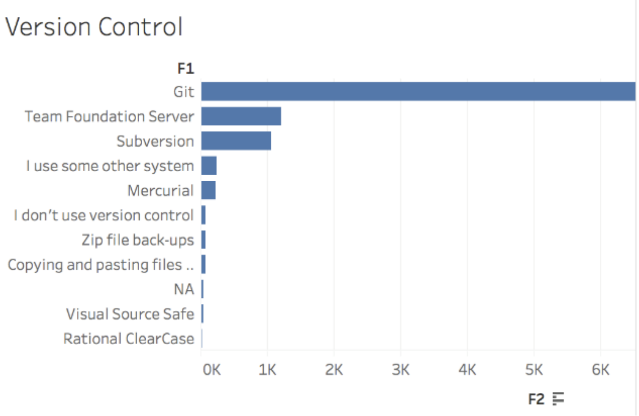

Figure 2: Popular version control

3) Comparison of employment status in 2017 and 2018

As we can see overall employment has been increased in 2018 and the percentage of full-time employment also increased in 2018.

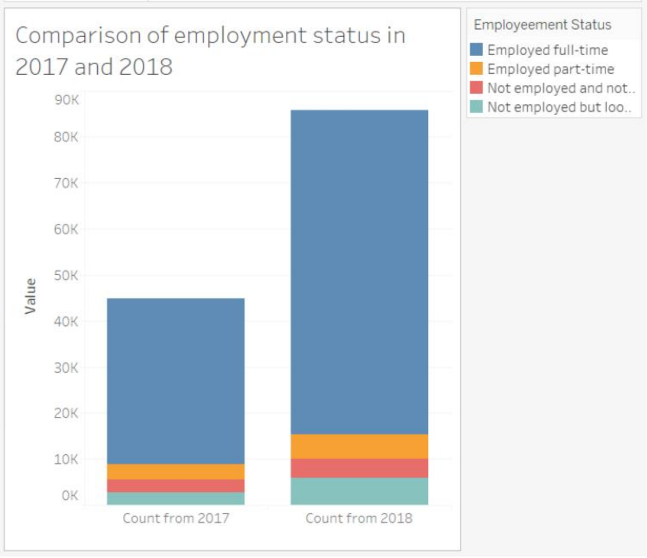

Figure 3: Employment comparison from 2017 and 2018

4) The popularity of Stack Overflow amongst age groups

Majority of users of this expert platform are from 25-34 years age group. Generally, in this age group peoples willing to work in software development and after 35 they might want to move managerial positions. Below report also highlights distribution on gender in which red shows males and females denoted by green color.

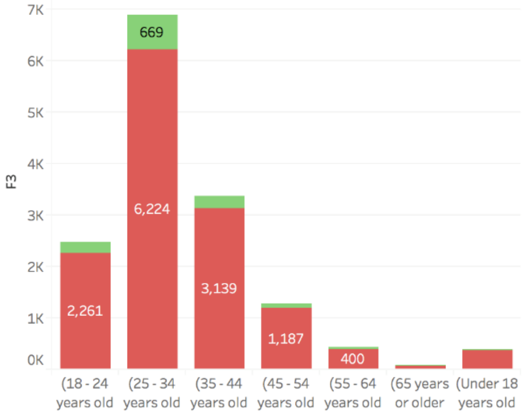

Figure 4: Stack overflow popularity

5) The motivation of developers for advanced AI

Advanced Artificial Intelligence is rapidly increasing in the software industry with pros and cons. From developers point of view "Increasing automation of jobs" is the main reason behind it as shown in below report.

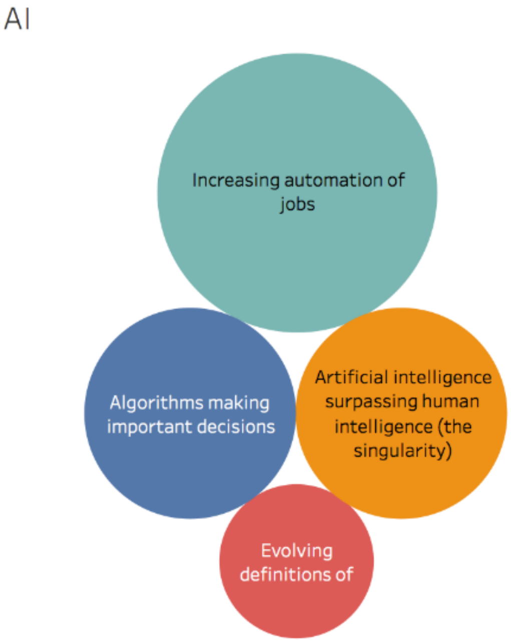

Figure 5: Motivation of developers for advanced AI

6) *Job Satisfaction of Full-stack Developer*

As we can see below results the majority of full-stack developers are moderately satisfied with their job profile. In summary, the majority of developers are satisfied for their role as full stack developers and those who are inclined towards dissatisfaction reasons may be different and which may include salary or employer and should be analyzed by the organization in which these dissatisfied peoples are working.

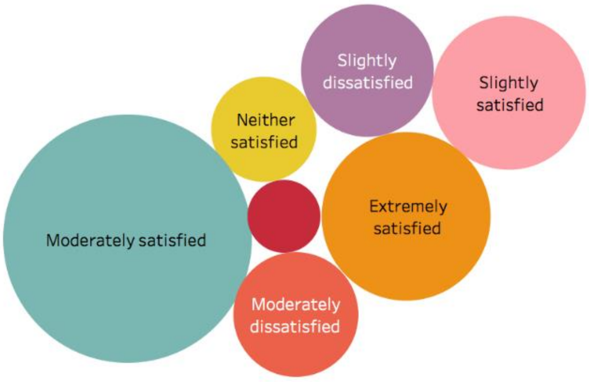

Figure 5: Job Satisfaction of Full-Stack Developer

V. Conclusion and Future Work

This study project done on Stack Overflow developers survey held in 2017 and 2018 which is very recent and includes opinions from over a million software professionals working across the globe with a different organization. As the survey was done by the privately held website and one of the major platforms for developer's community to get "Expert advice", this study keeps its own importance. Analytics highlighted in this study project is formed by honest opinions of over a million professionals without boundaries of organization, country and age group so we got the better results. Furthermore, despite the wide range of tools for data analytics "Hadoop Ecosystem" provides a rich set of frameworks such as Hive and Pig for analysis purpose which provides flexibility, scalability, fault tolerance and speed in big data processing. In this project, we used HDFS, Hive, Pig which are the backbone of the Hadoop Ecosystem and provides better results. For the visualization purpose, Tableau has been used which generates different types of graphs for classification.

In summary, this study project identified as the most trending tools and preferences of software developers towards technology. "C#, JavaScript and Java" hold the top positions in the trending programming languages and popularity of versioning tool is dominated by "Git". Employment status is seeming to be changing in 2018 and stack overflow is mostly popular in "25-34 years old" age group in the United States. Trending technologies such as advanced Artificial Intelligence are famous amongst software developers because of "increasing automation of job". This study is among, if not the first, addressing trending technologies and tools in the software industry can show the path to newcomers in the industry and gives the basic understanding of current trends and preferences of software developers. This study can be extended with different datasets and parameters.

References

[1] Kaggle.com. (2018). Stack Overflow 2018 Developer Survey | Kaggle. [online] Available at: https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey/home [Accessed 7 Aug. 2018].

[2] Insights.stackoverflow.com. (2018). Stack Overflow. [online] Available at: https://insights.stackoverflow.com/survey/ [Accessed 7 Aug. 2018].

[3] Ebert, C., Shankar, K., 2017. Industry Trends 2017. IEEE Software 34, 112–116. https://doi.org/10.1109/MS.2017.55

[4] Raunak, M.S., Binkley, D., 2017. Agile and other trends in software engineering, in 2017 IEEE 28th Annual Software Technology Conference (STC). Presented at the 2017 IEEE 28th Annual Software Technology Conference (STC), pp. 1–7. https://doi.org/10.1109/STC.2017.8234457

[5] Santos, R., Werner, C., Barbosa, O., Alves, C., 2012. Software Ecosystems: Trends and Impacts on Software Engineering, in 2012 26th Brazilian Symposium on Software Engineering. Presented at the 2012 26th Brazilian Symposium on Software Engineering, pp. 206–210. https://doi.org/10.1109/SBES.2012.24

[6] Ahmed, T., Srivastava, A., 2017. Understanding and evaluating the behavior of technical users. A study of developer interaction at StackOverflow. Hum. Cent. Comput. Inf. Sci. 7, 8. https://doi.org/10.1186/s13673-017-0091-8

[7] Tableau Software. (2018). Tableau Software. [online] Available at: https://www.tableau.com/ [Accessed 7 Aug. 2018].

[8] Pig.apache.org. (2018). Welcome to Apache Pig!. [online] Available at: https://pig.apache.org/ [Accessed 7 Aug. 2018].

[9] Hive.apache.org. (2018). Apache Hive TM. [online] Available at: https://hive.apache.org/ [Accessed 7 Aug. 2018].

[10] Miner, D. and Shook, A. (n.d.). MapReduce design patterns. 1st ed. O'Reilly Media, Inc.