Research On Software Testing Technology Under Big Data Background

Ma Hong
Hainan College of Economics and Business, Hai Nan

Abstract: With the development of Internet, the age of big data is coming. The new technological revolution brings new challenges to software quality and software test. This paper analyzes the development trend of big data age and the software feature and form changes come along with it. The author also analyzes the challenges of software test technology under big data background.

Key words: big data software test challenge

1.INTRODUCTION

In recent years, with Cloud technology and Internet of Things technologies becoming mature and pervasive, the data is in explosive growth. In the era of Internet, how fast does data generates? According to an investigation conducted by the Intel in 2013, the amount of transferring data per minute is 64,000Gb.The explosive growth of data arose the interest of industrial and academic circles. In 2008 at the tenth anniversary of the establishment of Google, NATURE devoted a special issue of the magazine to discuss the technical problems and challenges of Big Data. From 2007 IDC came up with the idea of Data Universe to describe data. According to IDC, the total amount of data is added up to 40ZB by 2020. How massive is 40ZB data? If the sand on the earth is about 70,005,000,000,000, the data on the earth is 57 times more than the sand on the earth. The application of data has a lot of successful examples including the circles of Internet, astronomy, atmosphere, genome, biology and chemistry, etc. According to a statistics conducted by Gartner, among 48 hottest new technologies big data ranks the first.

2.The Changes Of Software Under Big Data Background

The big data has the following four characteristics whose initial is V: First, Volume. The rank is from TB to PB and even EB. Second, Variety. It means the types of the data include a wide range of items, from web log, videos, photos, locations to texts. Third, Veracity. It means the data needs to be dig to find out relevant information. Fourth, Veloctiy. It means the instantaneous analysis of data, which is different from the traditional data dig technology. All the four characteristics describes the feature and form changes of software .

2.1Unfeasible Big Data Handling Techniques

The definition of big data in Wikipedia is that enormous information, most commonly in the form of a series of binary digits, stored on aphysical storage medium for manipulation by a computer program. The big data cannot be collected, searched or dealt into valuable information under our handling techniques. Sam Madden from American MIT makes further exposition that we call it big data for the handling techniques is out of our ability.

2.2Unknown Results of Big Data Software

Academician Li Guojie in “Research of big data problems” puts it, big data software is like hunting a fish in the ocean rather than looking for a needle in the ocean. We can look for the needle is because we know that there is a needle in the ocean. But the dig of big data is that we don’t know whether there is anything in the ocean. The best result is to find out the hidden relations between data and what they are. What’s worse, we don’t whether there is a result. Hunting a fish in the ocean cannot make sure that the fish we hunts slip away from the net or the fish we hunts is what we want.

2.3Unlike Thought Pattern to Traditional Software

In traditional software, we build mathematical models first. There are instructions in the software. The relationship between data is certain, especially the input and output of software, after which we use procedures to establish mathematical models. The purpose of researching big data is to find out the relations between data. The relations and features of data exist on the Internet in an uncertain way. We don't know the special way of existence and whether it exists. So the big data software has unlike thought pattern to traditional software.

3.The Problems Of Software Test Under Big Data Background

3.1 ORACLE Problem in Software Test

In 1945, a moth flew into Grace Hopper's Harvard Mark II computer, giving birth of the first computer error in history. It has been 70 years since the first computer error. From the perspective of test, several test ideas and strategies have been come up with. The most representative opinions are from Dr. Bill Hetzel(The Representative work is Complete Guide to Software Testing)and Glenford J. Myers (The Representative work is The Art of Software Testing). The former holds the view that test is to justify that software is working. The latter holds the opposite view. Glenford J. Myers thinks that test is to justify that software is not working. He puts it that we should focus on the errors of the software and find out as many errors as possible. Then we infer that the software is not work with reverse thinking. No matter what kind of concepts we hold, the premise is that we already know what kind of situation is right, what is wrong. This seemingly inevitable premise no longer exists under big data background. Test software requires us to compare the actual results with the expected results to get whether the software is working or not. This is ORACLE problem is software testing. But in a lot of circumstances under big data background the output can not be determined directly.

3.2 The Ability Comparison of Spear And Shield in Software Test

As two aspects of the problem, spear and shield is inevitably developing at the same time. In the C / S and B / S performance testing at early times, we use controller to coordinate and send a service request in order to finish stress test. The test load generators are based on the local physical host. IBM's Performance Tester, HP's Loadrunner and Mircofocus's SilkPerformance use this way to achieve software test. For application system constituted by a small number of servers that is hundreds and thousands, this approach can meet the needs of this application. While under big data background, in order to deal with the explosive growth of data, the number of data processing platforms is on PaaS platform that is dynamically extensible. Hadoop platform supported by the dund of Apache is currently the most famous big data processing Series. Data processing software can be used on platform with ten million servers. But the platform with such enormous servers is too difficult for us to build. The ability to test the client-side may not be able to meet the requirements of the server-side. Thus the ability of spear and shield is exactly reversed.

3.3 Validity Brings Correctness Decision Problem

One of the basic characteristics of big data is validity. Expected for structured data, unstructured text data, semi-structured social media data and location-based data are in rapid growth. Structured data is often stored in a relational database table RDBMS or structured file (XML). The structured data of a physical nature is simply exists in files and RDMBS table for the correctness of software test. Unstructured data doesn't have any format, which is stored in a document or web page.Semi-structured data doesn't have predefined formats. But the data structure can be deduced from the basic model. Using

unstructured data as software will be an extremely complex and time-consuming work. Although we can use some automation tools to extract the contents of unstructured and make it structured data, but the data itself may occur internal exception and data formats may change during the test. Data validity exists in the input side and output side. For this reason it is difficult to decide the correctness of the test software. If the output is a massive of unstructured data, the determination of the correctness will be more difficult.

3.4 Pesticide Problem in Software Test

Boris Beizer proposed the pesticide problem of software test in 1990.The so-called "pesticide" is used to describe the immune competence of the software. The more we do software test, the stronger the immune competence is. This phenomenon the same as using pesticide to kill insects, long-term use of a farming medicine can make the pest resistant to the medicine. The pesticide finally loses its effectiveness. In the developing process of all kinds of software, various errors are found in the early models. We verify or calibrate the components again and again. Without writing code these components has a natural immunity to the test. Pesticide problem asks the software testing technology to upgrade constantly in order to find the errors in software. In tests, we can find more errors in the beginning. The number of errors can be found in the post-test gradually slow down or even stop. With the significant effect of pesticide problem, we must find a way to stop the situation.

3.5 Positioning Problem in Software Test

At present, lots of researchers find the data dependency rather than data causality. In fact, finding causal relations outweigh the cost and difficulty than finding the correlation. What is data dependency? It is when A occurs, there is bound to be B. But the reason is due to C that A and B occurred or because A leads to B? Take the example of academician Li Guojie: Before the rain, it is common to swallows flying low. From chronological perspective, the relationship between the two may be complex. The reason for the swallow flying low is because of the rain. In fact the swallows flying low is the reason for the rain. Big data analyze and process all data rather than random samples. The most valuable data is abnormal data. The statistical data found out that all the new things come from abnormal data. The maximum value of abnormal data gives us the chance to discover our cognitive defects. The modeling data utilize the defects to discover unknown new things and improve our cognitive model. The test design systems or test evaluation systems is very important for us to dig the high quality bugs. If the original program uses all the data, is it possible to construct another equivalent data set? Are all of the input data equal to all input possibilities? Is it necessary that the test exists?

4. CONCLUSIONS

Big Data is coming so ferociously, bringing great changes in technological development, industrial applications and social management. In order to meet the age of big data, various countries have made great data plans, programs and policies to stay in the leading positions of big data industry. The most famous acts includes the European Commission Members' "Open Data Strategy of European Union", the US government's "Big Data Research and Development Initiative" program, the United Nations' "Global Pulse", the French government's "Digital Roadmap" and so on. In China, all regions have launched big data strategies. Guangdong and Shanxi provinces have taken the lead in big data industry. Faced with the challenges of big data, we advocate integration of the national evaluation agencies and build nationwide test resources. After more than a decade's development, various national agencies have formed basic data environment with its own features.

In order to deal with the challenges under the background of big data, the following perspectives need to be improving. First, we should try to explore for intelligent data

processing in order to deal with ORACLE problem. The design of test case includes the input data is not limited to border requirements, property requirements but the input of data distribution characteristic size, sample set, and an output ORACLE's evaluation. Second, the test technology shall transfer into multiple measurements from a single test technology. Multivariate testing technology integrated application is an effect way to avoid pesticide effect. Diversified testing technology includes two meanings: a known test integrated application test methods and traditional testing methods based on study new testing methods. Third, we should construct automatic cloud environment. The client capacity of the environment must match the needs of the server. The cloud itself especially IAAS platform cloud testing environment should bear the load generator and explore new requirements and new methods of monitoring.

References:

[1] Boris Beizer. Software testing techniques[M].2nd ed. Van Nostrand Reinhold Co. New York,NY,USA,1990.

[2] Jeffrey Dean, Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Communications of the ACM-50th anniversary issue: 1958-2008 CACM.2008,51(1): 107-113.

[3] Ming K Hu. Visual pattern recognition by moment invariants [J]. IRE Transactions on Information Theory, 1962, IT－8: 179-187.