IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Special Issue, 4, Feb 2019 E-ISSN: 2347-2693

Prediction of Data Ware House Model using Dynamic Function Point Analysis

K. Bhuvaneswari

Department of Computer Application, Idhaya College for Women, Kumbakonam, Tamilnadu, India

Corresponding Author: bonish88@gmail.com

Available online at: www.ijcseonline.org

Abstract— Approach for estimation of Data Ware House(DWH) Projects/Data Marts using Function Point Analysis is an ETL Development, Enterprise. The Objectives are the burden of maintaining the composite and enterprise data model by the data warehouse is not directly recognized. Often a development team's "hidden" efforts in delivering the support architecture of a data warehouse are compared unfavorably with more traditional, and highly visible, user functionality. Unlike other traditional data base systems, a Data Warehouse uses other software systems as data sources and does not create new information, which generally would be more static in nature. So, applying Function Point Analysis to Data warehouse applications became more tedious as Data warehousing itself has some peculiarities of its own compared to traditional OLTP(Online Transactional Processing) applications. Data Warehouse/Data Mart are a special type of applications, with particular characteristics such as the fact that the users only use the software system for queries and report generation and not for data update, the fact that development is based on existing data of other systems without generating new information, and the fact that it follows a different development process than the traditional OLTP software systems. It is necessary, therefore, to adapt (rather than to exactly follow) the size measurement approach defined for the traditional OLTP systems so that they consider the specific characteristics of Data warehouse/Data Mart and generate more accurate estimations. The proposed approach helps in estimating Data warehouse/Data Mart Projects using Function Point Analysis especially for ETL operations, in a more traditional and systemati c way.

Keywords— *FPA* - Function Point Analysis, OLAP - Online Analytical Processing, ETL - Extraction, Transformation, Loading, OLTP - Online Transactional Processing, DWH - Datawarehouse

I. INTRODUCTION

Du The Function Point Analysis measure, quantifies the functionality asked for and provided to the user, based on the user's requirements and high level logical design.The focus was on the unhidden, user identifiable, external aspects of software. The approach was to enumerate these aspects and combine these using weights to arrive at a composite functional measure of the softwaresize.The Function Point Analysis method that evolved was based on analysis of considerable past project data and use of trial and error to fit a model that seemed to best represent the software size in terms of the functionality.International Function Points User's Group (IFPUG), evolves the FPA method and periodically releases the Counting Practices Manual for consistent counting of function points across different organizations.Functional Point Analysis is a popular method for estimating and measuring the size of application software based on the functionality of the software from the user's point of view.

The following queries can be answered by using the proposed approach.Industry is concerned whether the estimates given for a project or group of projects is done in a logical and scientific way?Have the inputs collected in a consistent fashion? Does the evaluation criterion meet the industry standards?Without Standardized definitions, how it is defined a person-month or a defect or a line of code can be verysubjective.

The function point methodology is well defined and has been proven to yield a consistent resultwhen used by experiencedpractitioners.Coupled with other well-defined metrics, such as hours (instead of person-months), the resulting measure, such as hours per FP, can be used with greater assurance of an "apple to apple" comparison. FPs is also an effectivenormalizing agent.It represents the size of the deliverable. A FP count is not influenced by the technology or the languages beingused.

Scope:

In-Scope: Estimation mechanism for ETL operations, OLTP applications.

Out-scope: Anything other than ETLoperations
Exclusions: Reporting, Messaging, OLAP

Pros:

1. FPA is a scientific way of sizing the functionality of an application, and definitely it can be adapted to data Warehouse applications, with some modifications.
2. One can go deeper into the functionality with the help of FPA, and can find out the loopholes that exist as part of the requirements gathering and can caution the client upfront about them.
3. The approach that is designed for FPA for data warehousing applications, speciallyfor ETL covers comprehensively all the ingredients that are required for sizing theapplication.

Cons:

1. Still it is too early to say whether FPA is the right approach for Data ware housing/Data Mart type of applications as they are far different from traditional applications. During Pre-sales stage, as the functionality may not be well defined, it is yet to be seen, whether FPA can be used for Ball-park estimates ornot.
2. As most of the activities like monitoring a job, croning a job are yet to be put under some function in FPA, it is still not sure whether the enter activities are covered for various phases of the Data warehouse applications.
3. The Industry have not defined productivity rates for some of the most commonly used tools in Data warehousing, which is adrawback.

To better control the time, cost as well as resources assigned to software projects, organizations require a proper estimate of their size even before the projects actually start. Accordingly, a number of approaches were proposed to estimate the size of a software project, as the well-known Function Point Analysis (FPA), which is largely used in traditional software development projects. However, it is observed in software Industry that it is not fit for data mart software measurement. Data Mart (DM) systems have particularities in their development that are different from the traditional software systems (e.g. a DM uses other software systems as data sources and does not create new information). It is important, thus, to have a measurement approach that considers those particularities while measuring the DM size. In this paper, presented an adaptation of the FPA approach for DM sizemeasurement.

Measuring the size of a software system is an important issue for project management since it is used to set more realistic expectations for the user, to better know the software patrimony of the company, to get and to improve time, cost and resources estimations. It is essential that the size measurement be as close of the reality as possible, since it impacts other variables.

a). Most of these approaches aim at measuring the size of any type of software system, whatever the technology may be. However some authors disagree with this view and argue that each technology has specific particularities, which must be taken into account.
b). They also argue that approaches like the FPA were defined more than two decades ago considering hardware and software technologies that are largely obsolete today. It is totally agreed. For example, Data Warehouse/Data Mart are a special type of software, with particular characteristics such as the fact that the users only use the software system for queries and not for data update, the fact that development is based on existing data of other systems without generating new information, and the fact that it follows a different development process than the traditional softwaresystems.
c). It is necessary, therefore, to adapt the size measurement approach defined for traditional systems so that they consider the specific characteristics of Data Warehouse/Data Mart and generate more accurate estimations. In this Project it is proposed an adaptation of the FPA for Data Mart sizemeasurement.
d). The choice of this approach is due to the fact that the FPA is one of the most popular approaches of the market.

To define a measure for a software product, there are two main approaches that evaluate size, effort and time: the parametric approach, that estimate the software delivery size using a predictor that can be more easily determined earlier in the software lifecycle, called metric (e.g. line of code, FPA); and, the heuristic approach, that, based on historical comparisons, design the cost specifications and quick and-dirty estimates.

Estimating data mart size

- There are substantial differences among the construction of transactional software and the construction of a DataMart.
- The development process of the second involves a lot of treatment of data from software systems already in use; the final product does not provide updating functionalities to the end user but only is available for them to mine and query the data with appropriate tools; or the system needs to be continually updated with new information from the source systems. Also, the end user does not visualize some important development activities that impact the systemsize.

Because of those differences, the FPA is not completely adequate for Data Mart size measurement.

The reused characteristics

- As described before, characteristics originally proposed in the FPA approach may be used to estimate the size of Data Mart systems, they are: distributed data processing,

performance of the system, reusability at different stages and operationalease.

- Distributed data processing describes the degree to which the application transfers data among components of the application.
- The development process of a Data Mart prepares the data to be queried by the end user with data accesstools.

Scored this characteristic, as in the traditional FPA:

0. The application does not aid in the transfer of data or processing functions between components of the system.
1. The application prepares data for user to process on another component of thesystem.
2. Data is prepared intended for transfer, then is transferred and processed on another component of the system (not for end userprocessing).
3. Distributed processing and data transfer are online and in one directiononly.
4. Distributed processing and data transfer are online and in bothdirections.
5. Processing functions are dynamically performed on the most appropriate component of thesystem.

II. PERFORMANCE

*Performance*describes the degree to which response time and throughput performance considerations persuade the application development.

- It fits to the Data Mart that works by means of high volume of transactions; even if OLAP tools support thesetransactions.
- Depending on the kind of project the degrees 4 or5 would beindicated.

Scored this characteristic, as in the traditional FPA:

0. No special performance requirements were stated by theuser.
1. Performance and design requirements were stated and reviewed, except no special actions were required.
2. Response time or throughput is critical during peak hours. No particular design for CPU utilization was required.
3. Response time or throughput is critical during all business hours. No particular design for CPU utilization was required. Processing deadline requirements by interfacing systems areconstraining.
4. In addition, stated user performance requirements are stringent enough to require performance study tasks in the designphase.
5. In addition, performance analysis tools were used in the design, growth, and/or implementation phases to meet up the stated user performancerequirements.

III. REUSABILITY

Reusability describes the degree toward which the application and the code in the application have been specifically designed, developed, and supported to be usable in other applications.Considered reuse in Data Mart similar to traditional systems.Scored this characteristic, as in the traditional FPA:

0. No reusablecode
1. Reusable code is used within theapplication.
2. Less than 10% of the application considered more than one user'sneeds.
3. 10% of the application considered more than one user'sneeds.
4. The application was specifically packaged and/or documented to ease reuse, as well as the application is customizes by the user by source codelevel.
5. The application was specifically packaged and/or documented to ease reuse, and the application is customized for utilize by means of user parameter maintenance.

Operational ease describes the degree to which the application attends to operational aspects, such as startup, backup, and recovery processes.

- Fitted for Data Marts for the extractionand loading batchprocedures.

Scored this characteristic, as in the traditional FPA:

0. The user stated no special operational considerations other than the normal backup procedures.

1 – 4. Depending of the existence of different items in the application: effective start-up, backup, and recovery processes provided, with or not operator intervention required; and the minimization of the need for tape mounts or paper handling.

5. The application do not require any operation intervention other than start up or shut down.

The adapted characteristics

Two characteristics proposed in traditional FPA were considered relevant to Data Mart but were reinterpreted based on Data Marts features.

These characteristics are end user efficiency and complex processing.

- *End user effectiveness* is originally defined to describe the importance of human factors and ease of use for the applicationmeasured.
- It's adapted this characteristic to consider the degree to which data are aggregated in order to provide better performances for the end user queries.
- This is because with Data Mart, better performance means better efficiency to the end user.

Renamed this characteristic Amount of Aggregation, and score it as follows:

0. No aggregationidentified.
1. One level of aggregationidentified;
2. Two or three level of aggregationidentified.
3. For five level of aggregationidentified.
4. Six or seven level of aggregationidentified.
5. Eight or more level of aggregationidentified.

- *Complex processing* is originally defined to describe the degree to which processing logic influences the development of the application (such as mathematical processing, special auditing processing,etc.).
- These processing activities do not apply to Data Mart, but the number of transformation activities needed to build it can be seen as a complex processing.
- Therefore, adapted this characteristic to consider the number of transformation activities required. Those activities make the development process morecomplex.
- Renamed this characteristic Data Quality to better represent its new meaning (more transformation activities are required for poor qualitydata).

Scored this characteristic asfollows:

0. No transformation activityneeded;
1. one or two transformation activitiesneeded;
2. three or f transformation activitiesneeded;
3. five or six transformation activitiesneeded;
4. seven or eight transformation activitiesneeded;
5. all transformation activitiesneeded.

The new characteristics

Finally, and as described previously, eliminated eight characteristics originally defined in FPA because they never influence the Data Mart development process (null value).

These characteristics are: data communications, heavily used configuration, transaction rate, online data entry, online update, installation ease, multiples sites, and facilitate change.

- On the other hand, defined seven new characteristics to be integrated in the value adjustment factor(VAF).
- They are based on the difficulties one may encounter during the ETL (extraction, transformation and loading)process.

It is identified this process has the most critical and time-consuming one in the Data Mart development.

- According to Lokan and Abran, the characteristics for the VAF aim at covering different aspects of a software product that are not otherwise considered in the countingprocess.
- These aspects are: complexity, performance, user support, quality, architecture, interaction, constraints, interface, operation, reusability anddocumentation.
- To define the new characteristics, carefully analyzed the ETL process, reviewed the relevant literature, and interviewed experts to understand how these aspects may influence the development of a DataMart.
- Reusability, constraint, performance, architecture, as well as quality are already covered by the six characteristics reused oradapted.

The complexities along with operation of the extraction phase were covered by the definition of three characteristics: Amount of transactional system involved within the project, source data structure, also level of knowledge required for the developmentteam.These characteristics impact the extraction phase because, for example, the more source systems and type of source data structure are involved, the larger the project willbe.

- To cover the complexity and operation of the transformation phase, and also the user support and interaction, defined a newcharacteristic:

Frequency of Source Data Update

These aspects are also covered by the adapted characteristics.To cover the complexity and operation of the loading phase, defined a new characteristic: Volume of data (the bigger the final Data Mart the more difficult it is to load).Finally, defined two other characteristics that impact the whole process:

The use of appropriate tool for extraction and load, and the source systems documentation. The first one covers the complexity and operation aspects because it can ease the whole ETL process.

The second one covers the documentation aspect and is very important in the whole development process.For each one of the new characteristics defined a scale from 0 to 5 following the same convention used in traditional FPA.The scales were defined based on interview with experts and own experience.

Use of suitable tool for extraction and load, indicating the level of automation of the ETL process.

Score as follows:

0. Tools are used for 100% of the process of extraction and/orload;
1. tools are used for 80% of the extraction process and/orload;
2. tools are used for 60% of the extraction process and/orload;
3. tools are used for 40% of the extraction process and/orload;
4. tools are used for 20% of the extraction process and/orload;
5. notool used for the extraction process and/orload.

- Amount of transactional systems involved in the project describes the degree to which interfaces with other systems will influence the development of theapplication.
- When the amount of transactional systems is high,

it raises the complexity of the development, implantation and support of the application.

Score this characteristic as follows:
0. N/A;
1. the project involves 1 transactionalsystem;
2. the project involves 2 or 3 transactionalsystems;
3. the project involves 4 or 5 transactionalsystems;
4. the project involves 6 or 7 transactionalsystems;
5. theproject involves 8 or more transactional systems.

- Source systems documentation (existence of metadata in the source systems) describes the level of documentation of the source systems, to assist in identifying the metadata of the source data.

Scored this characteristic as follows:
0. All source systems havemetadata;
1. 90% of the source systems havemetadata;
2. 70% of the source systems havemetadata;
3. 50% of the source systems havemetadata;
4. 30% of the source systems havemetadata;
5. noneof the source systems havemetadata.

Frequency of source data update describes the frequency of modification of the transactional source systems resulting in constant alteration of the ETL process.

Scored this characteristic as follows:
0. No previsionexist;
1. 10% to 20% of the extraction/load filesmodified;
2. 20% to 30% of the extraction/load filesmodified;
3. 30% to 40% of the extraction/load filesmodified;
4. 40% to 50% of the extraction/load filesmodified;
5. more than 50% of the extraction/load files modified.

- Source data structure depends on the structure of the source data: VSAM, Relational (Ingress, Informix, Sybase), hierarchical(IDMS).

Scored it as follows:
0. N/A;
1. only 1 structure for the sourcedata;
2. 2 structures for the sourcedata;
3. 3 structures for the sourcedata;
4. 4 structures for the sourcedata;
5. more than 4 structures for the sourcedata.

- Volume of data that is a prevision of the volume of data of the project. The volume of data influences the size because one needs to guarantee good performances.

Scored it as follows:
1. low (till 500 gigabytes);
3. medium (from 500 gigabytes to 1 terabyte);
5. high (up to 1 terabyte).

- Level of knowledge required for the development team of the business rules of the transactional source systems. Influenced by the existence of ETL tool, because they compels the team to know all the business rules in order to appropriately define extractionforms.

Scored it as follows:
1. Little knowledge of the business rules;
3. average knowledge of the business rules;
5. high knowledge of the business rules.

IV. CONCLUSION

One of the major difficulties in project management is to estimate the project size to be able to deduce important factors like cost and effort. There are many approaches to estimate the size of a project. No one is better than the other in all situations. The approach must be chosen and adjusted depending on the particularities of the system that one intends to develop. Considering that Data Mart systems have substantial differences in their conception and way of development from the conventional systems, I have given an approach of size measurement (the well known Function Point Analysis) to this context.

One can say that applying FPA to a particular domain usually requires some interpretation of the FPA rules for that domain, what I proposed in this Project is an interpretation for Data Mart projects. I am aware that more applications of the proposed approach are necessary to better confirm its adequacy to Data Mart project measurement, but I believe this approach makes a considerable improvement in solving the problem of sizing Data Martsystems.

REFERENCES

[1]. INMON, W.H., Definition of a Data Warehouse.1999.
[2]. KIMBALL, R.,THORNTHWAITE, W.,REEVES, L..ROSS, M.,The Data Warehouse Lifecycle Toolkit.
[3]. New York: John Wiley & Sons, 1998. FENTON,N., PFLEEGER, S. Software Metrics A Rigorous & Practical Approach. Boston: PWS Publishing Company, 1997.
[4]. ISO/IEC 9126:2001.Software engineering Product quality.2001.
[5]. IFPUG. International Function Point Users Group. Function Point Counting Practices Manual: Release 4.1. Ohio: IFPUG. 2000.
[6]. Adapting function point analysis to estimate data mart size by Angelica ToffanoCalazans, MarcalDe Oliveira, RildoRibeiroDos Santos.

