The Open Data Interface (ODI) Framework for Public Utilization of Big Data

Hwa-Jong Kim
Dept. Computer Engineering
Kangwon National University
ChunCheon, Korea, 200-701
hjkim3@gmail.com

Seung-Teak Lee
IT Convergence Service Dept.
National Information society Agency
(NIA), Seoul, Korea
leest@nia.or.kr

Yi-Chul Kang
IT Convergence Service Dept.
National Information society Agency
(NIA), Seoul, Korea
kangyc@nia.or.kr

***Abstract*—In the paper, the Open Data Interface (ODI) framework for public utilization of Big Data was suggested. Comparing to conventional Web based open APIs which provide restricted access to the internal database of large data companies, the ODI provides multi-level access of data in public domain, ranging from copying plain files to extracting executive summary. Through the ODI, users can share raw data, intermediate mined data, or graphical reports, and can contribute to public utilization of Big Data. In the National Information society Agency (NIA) of Korea, the ODI scheme is considered as a public data infrastructure to support big and small companies for future data mash up business. Many companies in Korea in the field of telecommunications and web services are interested in developing a collaborative public data infrastructure with the guide of government. In the paper, we proposed an initial test infrastructure for the purpose.**

Keywords-big data；open API；open data interface； public data

I. INTRODUCTION

These days, Big Data is attracting much attention because of its potential benefits to find valuable information from plain data. Big Data services include data gathering, data analysis, data mining, recommendation, prediction, and reporting by using various data sources such as, sales data, social network service messages, location information, and any related documents.

However, together with the prospecting advantages of Big Data, some problems are also expected in the Big Data world:

- Monopolization of data by large data companies
- Digital divide in data accessibility
- Big data traffic due to redundant copies

As the Big Data service is growing, a few large *data companies* such as Google, Facebook, Twitter, Amazon, Yahoo, Naver, or Daum (big web portals in Korea) will have a good chance of gathering valuable data every day. The large data companies can make use of the big data for marketing and service improvement, which again gathers more valuable data from the users. This will give more severe digital divide in data access capabilities for individuals and small companies.

The Big Data service will eventually incur big traffics. The volume of data in the world will grow, and as far as many Big Data services are introduced, a transformed data set will also be generated by many companies, institutes, and individuals. This will also produce redundant data set which is almost same to the original data except that only a very small part is changed. The main bottleneck of Big Data may come from the telecommunication channel rather than memory or computing resources. The bandwidth of channel is always limited, and costs high.

In order to handle above problems, we suggest the Open Data Interface (ODI) framework for public utilization of Big Data. The ODI is an extended framework of the conventional Web based open application program interface (API), but providing more flexible and unconstrained access interface. With the ODI, users can access raw data file, intermediate mined data, or graphical reports. With the conventional Web based APIs, users can get data from sites, but cannot put his or her processed (or mined) data to the sites. In other words, API users cannot contribute to build a public data infrastructure with a more rich set of raw data or mined data. With the ODI, users can put related data, processed data or even mining algorithms to the ODI infrastructure.

In the paper, we briefly review related work, and describe the concepts of the ODI framework and its operations.

II. RELATED WORK

Many companies (or institutes) provide Web based open APIs for users to access database in the company. There are more than 6000 sets of APIs, including Twitter, YouTube, Facebook, Google Maps, Flickr, LinkedIn, etc. [1]. But the open APIs have the following limitations (see Figure 1):

- Open APIs provide limited access to data in size and types
- Open API requires professional programming skill
- the purpose of the open API is for the company, and activity history is accumulated in the company

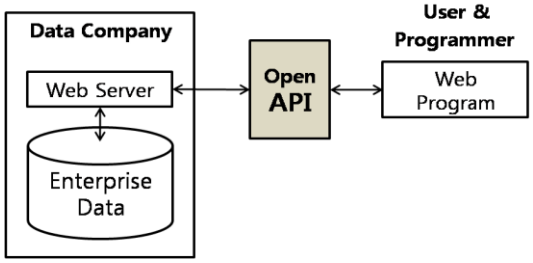

Figure 1. With conventional open API, the users should develop high-tech web programs to get data from the database of the company.

 ISBN: 978-1-61208-242-4

Even though the data company provides open API, it is usually limited in the scope and volume of the access data. Furthermore in order to use the open API, we need skilled programming. It is noted that the final purpose of the open API is for the company. It is not for public good.

Recently, it is suggested that Big Data should support improving public services in health, safety, and creating new business [2]. For this purpose, a public data space will be needed for easy sharing of data.

A related work is the Linked Data which was proposed to connect many types of data over the Web by using the Uniform Resource Identifiers (URIs), HTTP for identification, and Resource Description Framework (RDF) for contents description [3]. The Linking Open Data (LOD) project [4] uses the Web like a single global database to integrate data from heterogeneous sources. The LOD helps users navigate between related data sets through the semantic web. Some challenges of Big Data were investigated including data quality and lack of good use cases [5].

The Interaction design during the process of acquiring, analyzing, and using the Big Data is also becoming critical issue for success of Big Data [6]. Machine learning algorithms can be applied to large data sets over hadoop platform [7], and a cloud based prediction service is provided by Google [8].

III. DEFINITION OF THE ODI FRAMEWORK

A. Background of the ODI

The LOD was proposed to link related data over the Web, and therefore can be used for public Big Data service because Web is open to anyone. However, the LOD is mainly focused on connecting related data sets and finding them efficiently. But we need a more flexible framework which can integrate, besides the data itself, the machine learning algorithms used, and specific domain knowledge obtained from the case.

In the proposed ODI framework, we expect distributed contribution of users in processing (e.g., data mining) and utilizing the raw and intermediate mined data. The ODI provides a multi-level access and processing of information based on closely related data sets by many contributors.

B. Public Data Space (PDS)

The ODI framework model is shown in Figure 2, where the ODI is used to access the Public Data Space (PDS) by many contributors from government, enterprise and individuals.

The PDS is composed of Data Core and Contributed Local Data. Data Core is a marketplace where every data can be searched and accessed and processed. It includes some core public data. Contributed Local Data is data residing in the users' server, such as government (public data), enterprise (open data for the public), and individuals (privately processed public data). The PDS is a platform of sharing data, and can be regarded as a collaborative data warehouse.

Each user may have its own private data that is not shared with others.

C. Data Core

Data Core is composed of physical storage of data, virtual collection of data which is physically located in the user servers, and data analytics functions. Data Core also provides programmable platform which include gathering, analysis, and reporting functions. In other words, Data Core is composed of data and functions contributed by any users.

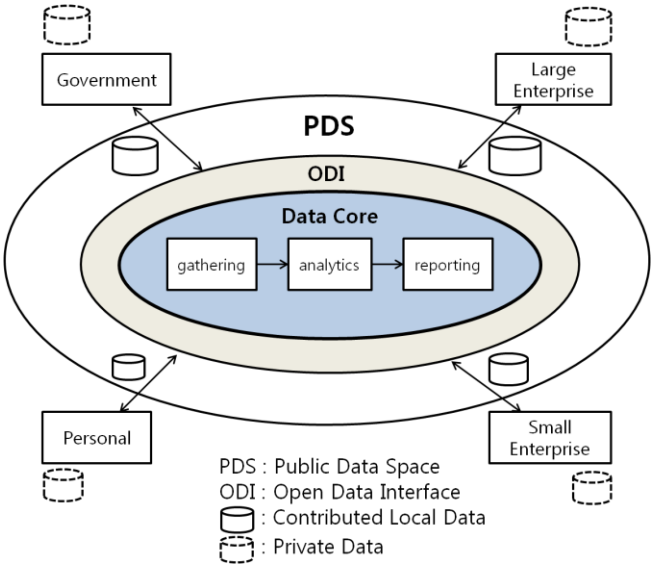

Figure 2. The ODI framework provides gathering, analytics, and reporting operations to the PDS

D. Open Data Interface (ODI)

The ODI uses the Data Core (its data and function) in order to provide various types of interface for gathering, analytics, and reporting of data. The goal of ODI is to provide easy but standardized interface in accessing data and sharing domain knowledge. We will explain the operations of the ODI framework.

IV. OPERATIONS OF THE ODI

A. Multi-level accessing

The ODI provides multi-level accessing to the PDS. The ODI provides various types of APIs to access raw data, processed data, abstract data, and they also do some operations of gathering, analyzing, and reporting. In other words, the ODI provides plain file copying, running machine learning algorithms, executive summary, or a graphic processing (see Figure 3).

In the ODI model, expert programmers are involved in developing the APIs in the Data Core and interface libraries in the ODI shell. Plain users may just input commands to the ODI to get some data or result.

 ISBN: 978-1-61208-242-4

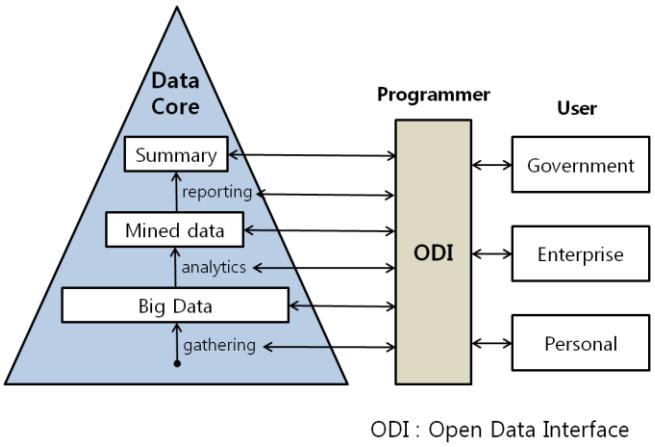

Figure 3. With ODI, the users need not to run web programs, but can get multi-level access to data and functions on the PDS.

B. *Functional Requirement of the ODI*

The ODI provides simple and standardized access to the Data Core. The Data Core is composed of data and functions operating on the data. The ODI should categorize the levels of accessing, i.e., levels of abstraction and processing levels. There are two types of levels:

- Data level (e.g., raw, processed, summarized data)
- Functional level (e.g., gathering, analysis, reporting)

The ODI interprets the response from Data Core to the users, and separates the role of user from being a professional programmer

V. USE OF THE ODI

Many companies and governments are gathering huge and valuable data in their domain every day, but do not fully utilize the data. It is because the data is isolated. For example, telecommunication companies may need banking information or Internet search history of their users for more intelligent services. In the future, new disruptive services will come from data mash up among various types of companies and government's public data. The ODI will help the extension of data mash up.

For data mash up cooperation among companies, we need an open infrastructure where each company can give (put) and get beneficial data in a standard and safe way. They want to sell processed (or screened) data and buy their missing data in an open market. The government should help the operations of the market through standardization of data format and access rules. We also need regulations in privacy-preserving data mining.

With the ODI, users can share their domain knowledge in the form of mined data or a new algorithm. For example, we can get top 20 news from a news portal. If an expert classified the top 20 news in an interesting way, he or she can share the idea by putting back the processed data to the news portal with the newly developed APIs. With this processed (or mined) data, other users may save time or memory by avoiding the same analysis.

We also expect the ODI may alleviate the drastic increase data traffic due to big data applications. The ODI will minimize the redundant copy of similar data by redundant (or similarity) checking. Traditionally, the usefulness of data mainly depends on the correctness of data. But in Big Data, the usefulness of data will mainly depend on timely access of data because the data preparation takes long time. For example, if two unstructured data set (e.g. blog data or news data) differs only by 1%, they may be regarded as a same data, and does not need to transfer them again for perfect coincidence.

VI. CONCLUSION AND FUTURE WORK

In the paper, we introduced the ODI framework which can be used to for public Big Data application by providing nationwide PDS infrastructure. The PDS and ODI framework need to be installed and operated by the government for public good, and minimizing digital divide in the coming Big Data era. The ODI framework is for easy access of data, algorithms and sharing success cases. It provides a kind of public data warehousing with related machine learning algorithms proven to be useful for some applications.

In the ODI model, the quality of data is not measured by the accuracy but by the usefulness of the data, which is evaluated by the users. We also hope that the ODI is used by individuals or small companies who want to create a new business in the Big Data world.

ACKNOWLEDGMENT

This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0301-12-1004)

REFERENCES

[1] http://www.programmableweb.com/ [retrieved: June, 2012]

[2] Alex Howard, Data for the Public Good, O'Reilly Media, 2012.

[3] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far" International Journal on Semantic Web & Information Systems, Vol. 5, Iss. 3, pp. 1-22, 2009.

[4] http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[5] Christian Bizer, Peter Boncz, Michael L. Brodie, Orri Erling, "The meaningful use of big data: four perspectives -- four challenges", ACM SIGMOD Record, Vol. 40, Iss. 4, pp. 56-60, 2012.

[6] Danyel Fisher, Rob DeLine, Mary Czerwinski, Steven Drucker, "Interactions with big data analytics", Interactions, Vol. 19, Iss. 3, pp. 50-59, 2012.

[7] Sean Owen, Robin Anil, Ted Dunning, and Ellen Friedman, Mahout in Action, Manning, 2011.

[8] https://developers.google.com/prediction/ [retrieved: June, 2012]

 ISBN: 978-1-61208-242-4