# A Systematical Scheme of Composite Analysis on Big Sensor-Data of Engineering Inspection

Min-Hwan Ok and Hyun-seung Jung

Korea Railroad Research Institute, Woram, Uiwang, Gyeonggi, Korea
mhok@krri.re.kr

**Abstract.** Dependencies, associations, correlations, and co-variances found out during big data analysis could unveil the basis of phenomena hard to understand. Recent paradigm of big data analysis has proven its potentiality with big data arose from social activities. Such big data could be generated in some engineering areas, since many kinds of sensors are equipped for researches in engineering. This work presents a scheme of an analysis against big sensor-data in a case of data measured on the railroad. The scheme is composed of procedurees for composite analysis comprised of engineering analyses and big data analysis. A role-based system diagram digests this data-intensive computing of the composite analysis.

**Keywords:** Big Sensor-Data, Composite Analysis, Data-Intensive Computing, Cloud Computing, Measured Data on the Railroad.

## 1 Overview

Advancement of instruments raises levels of data volumes in the scientific and engineering fields as higher resolution or rates found more precise or accurate analyses. Consolidated computing resources such as Cloud are also concentrating at processing the data volume larger than ever. Distributed computing resources are managed in the way efficiently allotting resources participating in the same computation considering how they are consolidated into one Cloud.

Sensor data collected by engineering inspection amounts in *Terabytes*, and the data types are mostly in one of three types: (sound) signal, (picture) image or video. In the case the data volume of terabytes is constituted with the data types in signal and image, generated by dozens of sensors, those data are heavy to be processed by an ordinary computing system, such as a single server system. Further if those data are collected against one physical entity for investigation into partitions of the entity, associations and correlations among the data could be revealed. This work schematizes a composite analysis on sensor data measured on the railroad, and stored in the Cloud.

Composite analysis comprises traditional engineering analyses with respective techniques and big data analysis. In the former engineering analyses, a research engineer analyzes the sensor data of a respective partition of the entity, the railroad in this

M. Matera and G. Rossi (Eds.): MobiWIS 2013 Workshops, CCIS 183, pp. 66–71, 2013.

work. The analysis result is input as a context about the partition and some domain parameters are derived from the context. The latter big data analysis involves these parameters, and the context enlarges domain knowledge in the Cloud.

In engineering inspection of railroad subsystem, traditional engineering analyses have focused on exact expectations for the partitions under mechanical engineering and those under civil engineering. The composite analysis would furnish associations and correlations between partitions under mechanical and civil engineering by further analysis with renewed techniques against big data.

## 2 Geographical Gathering Procedure of Big Sensor-Data

The data collected from physical world has errors with the sensor employed. In the work of a Worldwide sensor Web[1], the original data is trimmed by abstractions such as table, function, user-defined functions, or model-based views. Data is also subject to noise, and thus data filtering/cleansing is the first phase in the big data analysis.

Before collecting sensor data of the entity from the physical world, the data model should be established for the entity. Data collected from partitions are managed in accordance with their dependencies in the entity. Fig. 1 illustrates the data gathering procedure.

1. Collecting raw data into Data Bunches

Sensor data are measured and collected at the inspection vehicle with inspection instruments. A simple filtering is applied to the collected data and the filtered data is preserved in the moving storage.

2. Copy data bunches to Local Storage

The preserved data of the moving storage are retrieved and copied to a local storage of the Cloud. The retrieved data are copied through simplified classification and extraction.

3. Transferring data bunches to Global Storage

The copied data are filtered and cleansed in the local storage. The data are transferred to the big data storage before their computations.

Once transferred to the Cloud, the data are analyzed with respective techniques in engineering. Since a number of research engineers conduct engineering analyses on respective partitions, a common ontology against the entity is required for semantics shared among research engineers. A semantic Web technology[2] is adopted for provenance of each engineering analysis, in the ontology construction of the entity, for the identification of objects, and in the management of the engineer group.

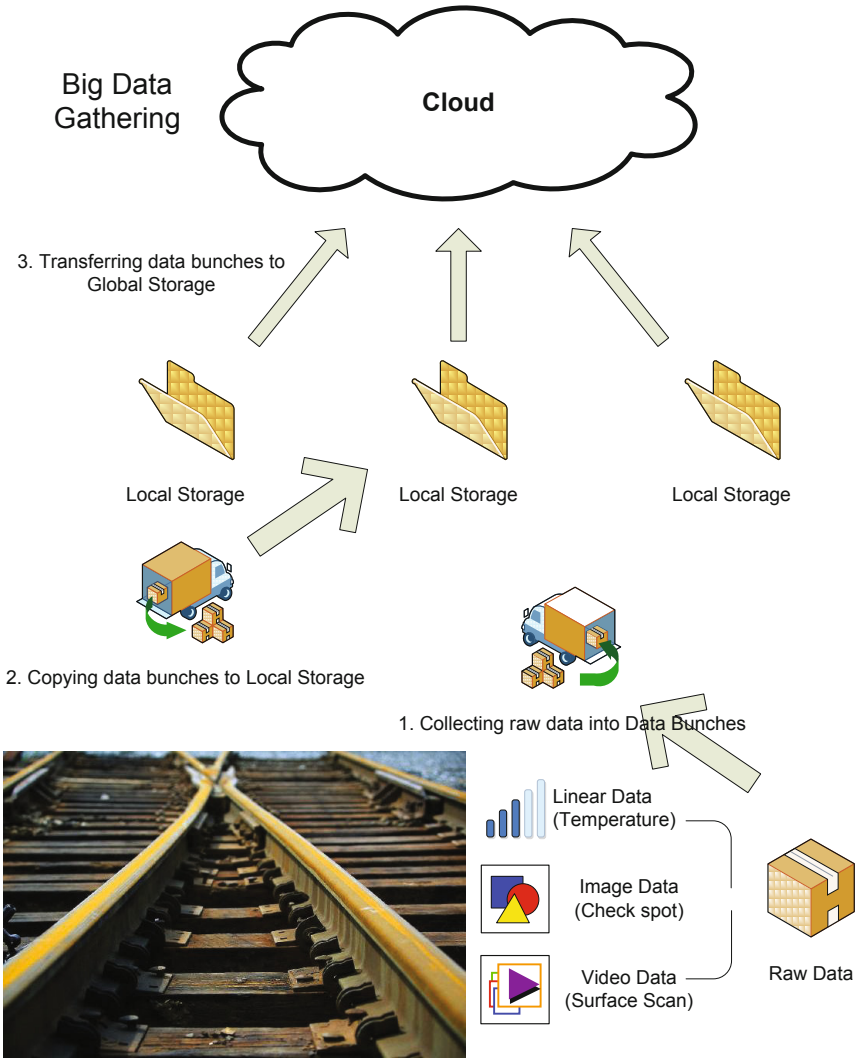


**Fig. 1.** The collected sensor data are accumulated in local storages and transferred to the Cloud

## 3 Composite Analysis Procedure on Big Sensor-Data

The traditional engineering analysis is conducted by individual research engineer appropriate to the partition of interest, and the analysis result is input as a context. These contexts could be used for the contextual search after the big data analysis, as those who search for data adequate to the query of complex semantics are research engineers. SOCRADES[3] integration architecture has a similar objective to our ones except that it provides functionalities of embedded devices than sensor data collected out of parts of the partition.

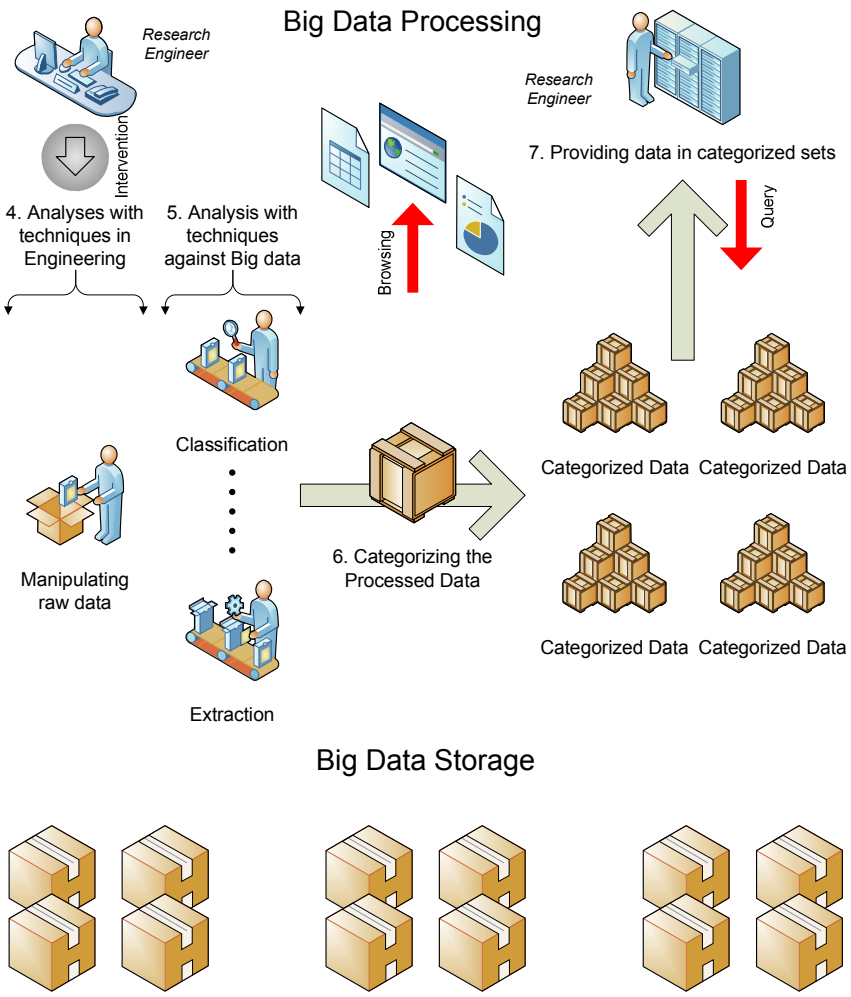


**Fig. 2.** After the composite analysis the processed data are provided in categorized sets

When the data transferred to the big data storage for their computations, they are duplicated for computation efficiency. The whole data is aggregated with local storages, and only processed results reside on the big data storage. Fig. 2 illustrates the composite analysis procedure.

4. Analyses with techniques in Engineering

Individual research engineer downloads the datasets of a partition and receives notes on inspection conditions(method of inspection, resolution and rate, environment description, and etc.). The person analyzes the datasets with proprietary analysis tools and save the analysis result so that the result could be included in the big data analysis.

5. Analysis with techniques against Big Data

Associations are analyzed with structured/unstructured data and correlations are analyzed from the associations. Searches into data could be parallelized with metadata produced by classification of data along the heterogeneity and extraction of metadata from the classified data.

6. Categorizing the Processed Data

An index is constructed with semantics gained throughout composite analysis. Relevant data are connected each other so that the connected data would form virtual datasets according to aspects of interest.

The data copied to local storages are accumulated for years and the data volume overwhelms the processing speed of an ordinary computing system[4]. Since the identical data set could be used in multiple calculations in the same time, correspondent dataset would be duplicated for the duration of its computation. Duplication strategy should be deliberated for the Cloud.

## 4 System Incorporation in the Cloud

In the system model, the system consists of local storages, big data storage, the big data processing and the information visualization. Local storages possess accumulated sensor data(raw data) and big data storage manages transferred sensor data for computations. These data are transient and those data processed by composite analysis remains in the big data storage. Big data processing is where the composite analysis proceeds and information visualization yields analysis reports on demand. Fig. 3 depicts the role based system diagram.

For the engineering analysis, the individual research engineer downloads the sensor data(raw data) from the big data storage through FTP. Later, graphical reports of the composite analysis is displayed for other research engineer through Web, and the processed data are provided in sets categorized after the composite analysis. In the engineering analyses, research engineers book respective partitions of the entity. Then the datasets of the partition are transferred to the big data storage and the research engineer is notified. Then these datasets are analyzed with respective techniques in engineering. This process resembles the book-loan one, and the framework in an IOT fashion[5] is adopted in the process of the engineering analysis.

The big data storage manages sensor data transferred from local storages. The transferred data are duplicated for computation efficiency. While duplicating the data, files are treated not to spread widely on many storage devices. The storage manager employs an advanced DHT-based distributed file system, D2[6] for this purpose. The feature of narrowly spread files has an additional benefit in system maintenance.

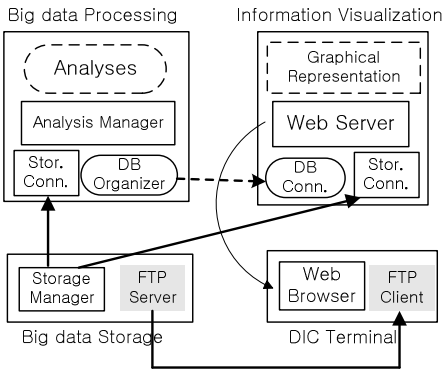


**Fig. 3.** System Diagram of DIC: The DIC terminal shows graphical reports of the analysis

## 5 Summary

In the composite analysis, big data analysis is influenced by domain knowledge enlarged with the contexts of each engineering analysis result. The graphical report of the composite analysis is displayed for other research engineer. Therefore the information visualization of the graphical report is of another importance in this work. We consider *Infographics* for information visualization, which is still undergoing. An inforgraphics library would be built in proportion to the number of categories. Visual analytics could be applied in big data analysis for information visualization.

For deep insight through big data analysis, some domain parameters derived from the context could be other key in big data analysis in addition to extracted metadata. The way the engineering analysis result would be input as a context demands a new discussion and it would be discussed in the future work.

## References

1. Balazinska, M., Deshpande, A., Franklin, M.J., Gibbons, P.B., Gray, J., Nath, S., Hansen, M., Liebhold, M., Szalay, A., Tao, V.: Data Management in the Worldwide Sensor Web. IEEE Pervasive Computing 6(2), 30–40 (2007)
2. Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D., Greenwood, M.: Using Semantic Web Technologies for Representing E-science Provenance. In: McIlraith, S.A., Plexousakis, D., van Harmelen, F. (eds.) ISWC 2004. LNCS, vol. 3298, pp. 92–106. Springer, Heidelberg (2004)
3. Guinard, D., Trifa, V., Karnouskos, S., Spiess, P., Savio, D.: Interacting with the SOA-Based Internet of Things: Discovery, Query, Selection, and On-Demand Provisioning of Web Services. IEEE Transactions on Services Computing 3(3), 223–235 (2010)
4. Jacobs, A.: The Pathologies of Big Data. Communications of the ACM 52(8), 36–44 (2009)
5. Giner, P., Cetina, C., Fons, J., Pelechano, V.: Developing Mobile Workflow Support in the Internet of Things. IEEE Pervasive Computing 9(2), 18–26 (2010)
6. Pang, J., Gibbons, P.B., Kaminsky, M., Seshan, S., Yu, H.: Defragmenting DHT-based Distributed File Systems. In: Proc. International Conference on Distributed Computing Systems, p. 14. IEEE (2007)