Big Data Challenges in Railway Engineering

Nii Attoh-Okine
Department of Civil and Environmental Engineering
University of Delaware
Newark, Delaware
Email: okine@udel.edu

Abstract—As Big Data becomes part of railroad data analysis, there are many challenges which need to be addressed by the railway industry. This extended abstract highlights some of the challenges from specific examples in railway engineering. The examples provided in this extended abstract cover both the engineering and the management of railroad applications.

Key words: Big Data; Railway; Rail Defects; Geometry Defects; Privacy.

I. INTRODUCTION

There are many opportunities to improve the productivity, reliability, velocity and safety of railroads. A major challenge is utilizing ‘Big Data’ to improve the factors [1]. Thomas [1] noted that with the smart infrastructure, North America’s railroads are working to capture ‘Big Data’ and figure out the best ways to utilize this data in various aspects of railroad engineering. Currently, railroads collect enormous quantities of data through GPS, AEI readers, electronic data exchange, video inspections, hand-held field tables, and many other sources. Furthermore, the data is growing both in quantity and quality and are more precise and frequent. Data of extremely large sizes are difficult to analyze using traditional approaches since they may usually exceed the limits of a typical spreadsheet. There are a few examples of ‘Big Data’ applications already in railway engineering. For example, Union Pacific Railroad is using Big Data approaches to improve operations in various ways including reducing train derailments, increased velocity of shipments and the reduction of emissions.

Henry [2] highlights the most salient applications of Big Data in urban rail transit, which ranges from urban planning activities with computerized processing of massive amounts of geographic and demographic data, complex signaling and train dispatch control, communication, and train tracking systems and ticketing information. Some examples of Big Data Analytics in urban rail transit systems include: a) Bay Area Rapid Transit which uses Big Data Analytics to monitor and analyze train arrivals and departures precisely, monitor flow of passengers based on ticket information and also for accurate projection of service schedules. Agencies which make use of such analysis include SEPTA Regional Rail in Philadelphia among others [3].

The development of advanced sensor and information technology in railway infrastructure monitoring and control has provided a platform for the expansive growth of data. This has created a new paradigm in processing, storing, streaming and visualization of data and information. The changes in technology include the possibility of installing sensors and smart chips in critical infrastructure to measure system performance, physical and other indicators of imminent failures. Furthermore, many of the railway infrastructure components have communication capabilities which allow data to be uploaded on demand.

An emerging terminology, Big Data is an important paradigm that should be considered in all aspects of railroad applications. According to a recent publication by IEEE [4], Big Data is a term with no set definition, as a result of rapid advancements in information technology. For example a decade ago, Big Data was measured in terabytes, now it is in petabytes. IEEE [4] noted the broader definition of Big Data is data that is too large and complex to be handled by traditional databases. Big Data can be formally defined as a collection of very huge data sets from which it is practically impossible to analyze and draw inferences. As stated earlier, rapid technological advancements have led to an exponential increase in the size of data.

Big Data usually has a multidimensional structure and can be characterized by the 5V’s:

- **Volume** - huge amount of data that needs to be processed, stored and analyzed. There is a need to develop algorithms that are scalable.
- **Velocity** - the indication of how to be analyzed. Some challenges include processing the data near-real time, or virtually real time.
- **Variety** - this relates to the different types of structured or unstructured data. Therefore the integration across different types of data is paramount.
- **Veracity** - this is the indication of integrity of the data. Therefore a robust and predictive algorithm capable of handling noise and incomplete data and information for decision making is required.
- **Value** - this is an important feature in Big Data. Mainly refers to the worth of the information being managed.

Fig. 1 shows the characteristics of Big Data. The data spectrum in railroad applications spans from large scale to the smallest possible scale [5]. The scales include:
TABLE I. TYPES OF BIG DATA

<table>
<thead>
<tr>
<th>Type</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Micro-data</td>
<td>The least aggregated level of data in very large data sets.</td>
</tr>
<tr>
<td>Meso-data</td>
<td>The mid-level of data aggregation in very large data sets, resulting from the collection of data that are a level-up from micro-data in terms of the kinds of information that is captured.</td>
</tr>
<tr>
<td>Macro-data</td>
<td>The most aggregated level of data in large data sets describes regional or geographic area.</td>
</tr>
</tbody>
</table>

a) Micro-data b) Meso-data, and; c) Macro-data. See Table I.

II. CHALLENGES

Jagadish et al. [6] presented a review of articles discussing the challenges of Big Data; most of the ideas presented in this article will be more tailored to railway engineering and other related areas. Also Chen and Zhang [7] highlighted challenges, techniques and the technologies for Big Data. Fig. 2. shows the Big Data analysis pipeline as presented by Jagadish et al. [6]. The figure, based on the feedback loops and characteristics, presents clear challenges encountered in Big Data analysis. Jagadish et al. [6] named 5 stages in the Big Data pipeline.

- Data acquisition
- Information extraction and retrieval
- Data integration, aggregation, and representation
- Modeling and analysis
- Interpretation

See Fig. 2. the stages and phases of the analysis connect directly to the challenges which are as follows:

1) Heterogeneity
Most of the data collected especially during track monitoring, both rail defect and geometry data, and also tonnage data are not homogeneous. In some cases, frequency data are used to do analysis, with little attention to actual values. The challenge in this instance is to develop an algorithm to transform the data so appropriate interactive analysis can be pursued, for example rail and geometry defect interactions.

2) Inconsistency and Incompleteness
It is very clear that there are uncertainty, errors and missing values for most of the data collected through, for example, video inspections. When data are collected from different spatial and geographical locations to make a general conclusion about track condition in part of the network, this can present another challenge on its own. This is because it involves the use of data from different databases which may be incomplete and inconsistent. There are even datasets with only subjective information. Methods for handling both objective and subjective information need to be introduced in these types of analysis. Currently, most of the data analysis in railroad applications focuses heavily on the objective side. There are few analyses which attempt to discuss the subjective information encountered during data collection. Unfortunately, analysis involving both objective and subjective data is missing in the Big Data railroad application literature.

3) Merging Data
Analyzing Big Data in railroads is not only about large databases but also the merging of different databases to extract information for further analysis. For example, how can one effectively merge ballast condition data and track geometry data or vice-versa? Also how can tonnage database be merged with rail defect database? All these are issues requiring a standardized approach and a systematic algorithm which can pass any statistical stochastic testing.
4) Timelines
Collecting track data in real time for maintenance purposes may have a time limit, therefore it is very important that real-time data analytic techniques should be implemented in railroad applications. As a result, timely maintenance can be implemented efficiently and effectively. For example, Union Pacific has developed alert systems to measure emissions every day based on 20 million daily pattern matches.

5) Privacy and data ownership
The use of electronic tickets and hence electronic information with the aim of improving services may bring some privacy issues to the fore. There is potential that personal transit ride data of an individual can be linked to both financial and health information. Currently, there is limited research addressing privacy issues, but none has been carried out in the area of passenger information.

These are a few challenges more specific to the railroad applications. Other issues concerning data capture and storage, data transmission, data analysis and data visualization have been addressed in the general Big Data literature.

III. CONCLUDING REMARKS

Big Data is the next frontier for both scientific and technical innovation in different parts of our society. Although big data is at the nascent stage, for railroad data both the quality and quantity are going to increase and this is the time for industries to make use of the available information and literature to develop industry-specific large data analytics approaches and methods. Successful completion of this agenda will put the railway industries in a new league through the development of advanced decision making tools for improving safety and efficiency for railroads.

REFERENCES