Assessing the Feasibility of Moving the SHRP 2 NDS Dataset

Prepared by:

Jeffrey A. Spotts, Staff Consultant to the Transportation Research Board of the National Academies of Science, Engineering and Medicine

Prepared for:

TRB Staff and the SHRP 2 Project SD-01 Safety Data Oversight Committee

September 1, 2019

Table of Contents

1. Executive Summary

The Naturalistic Driving Study (NDS) conducted under the second Strategic Highway Research Program (SHRP 2) generated a dataset comprising more than three petabytes (PB) of data. It is currently housed and maintained by the Virginia Tech Transportation Institute (VTTI) under a Cooperative Research Agreement with the Transportation Research Board (TRB), which is funded through March 31, 2020.

The NDS dataset represents a valuable resource to the research community that should remain accessible for a period of 20 or more years. As part of an effort to ensure the sustainability of the NDS dataset beyond 2020, the technical and economic feasibility of moving the NDS dataset to another custodian was assessed, which is the subject of this report.

The assessment followed a classic discovery process, which was intended to provide order of magnitude estimates of cost and effort to guide strategic decisionmaking.

The findings show that the NDS dataset is not a simple, static collection of data. It is large, complex, and dynamic, since it continues to be enriched by ongoing data reduction efforts. The data span multiple platforms. Data, tools, and processes involved are inextricably linked, and depend upon intellectual property developed by VTTI outside of the SHRP 2 program.

Migrating the SHRP2 NDS dataset is not a simple task of copying data from point A to B. The effort and costs involved are considerable, and would constitute a multi-year, multi-million-dollar undertaking. While it might be technically feasible to move the NDS dataset, considering all of the complexities, costs, time, potential disruption to users, and unknowns discussed in more detail in the body of this report, it is not economically feasible to consider this a viable option given the ongoing use of the data by qualified researchers and the budgetary resources available to support them at this time.

2. Background

The Naturalistic Driving Study conducted under the second Strategic Highway Research Program was a long-term effort to collect data to better understand how drivers interact with, and adapt to, the driving environment. It was expected that these data could offer the highway safety community an unprecedented opportunity to directly study driver behavior and its relationship with safety, representing a valuable resource that should remain accessible to the research community for a period of 20 or more years.

The NDS study involved the installation of sensors, video cameras, and other instruments on the cars of more than 3,500 volunteer drivers at study sites in Florida, Indiana, New York, North Carolina, Pennsylvania, and Washington. The study encompassed approximately 5 million trips and 30 million miles of normal driving. The resulting dataset comprises more than three petabytes of data, including internal and external video and sensor data, crash data, system data, participant demographic and assessment data, vehicle inventory data, and analysis data.

Following the conclusion of the data collection phase of the study, NDS data were made available to researchers worldwide. Because the study involved human subjects research, open access to the data is not feasible. Data use licenses (DUL) are required to use the data, any data involving personally identifiable information (PII) can only be viewed in a secure data enclave (SDE), and all original data must eventually be destroyed.

Because of the complexity of the data and restrictions on access to it, expert assistance is required to identify and extract subsets of data that researchers can use.

The NDS dataset is currently housed and maintained by the Virginia Tech Transportation Institute under a Cooperative Research Agreement with the Transportation Research Board, which is funded through March 31, 2020.

The author of this report, whose background includes experience in large scale data management and the long-term preservation of digital assets, was engaged by TRB to advise staff and the Safety Data Oversight Committee (SDOC) on information technology (IT) matters related to the sustainability of the NDS dataset beyond 2020.

3. Document Scope and Contents

This report is an assessment of the technical and economic feasibility of moving the NDS dataset to another custodian.

The remaining sections of this document are as follows:

Analysis Approach – Provides an overview of how the assessment was conducted.

Key Findings – Characterizes the major components of the dataset and the IT environment that supports it, the major functions involved in making useful and accessible to the research community, and the various issues and challenges that would have to be addressed in any migration scenario.

Conclusions – Presents a summary assessment of the feasibility of migrating the NDS dataset.

4. Assessment Approach

This assessment followed a classic discovery process. It involved site visits to VTTI, interviews with staff responsible for information technology, data reduction, data analysis, governance and compliance; demonstrations of software tools, observation of various business processes, and a survey of the compute, storage, and network resources required to support the NDS dataset.

The intent was to develop a sufficient understanding of the environment in order to:

- Characterize all of the components the NDS dataset, including hardware, software, networking, intellectual property, people, and processes.
- Identify the major issues and challenges involved in a potential migration.
- Roughly estimate the time and cost of such an effort.

Note that this assessment was not intended to provide a detailed project plan or a precise estimate of time and cost of migration. That would require an extensive, bottom-up evaluation by a team of subject matter experts. Rather, the intent was to provide order of magnitude estimates to guide strategic decisionmaking.

5. Key Findings

The NDS dataset is not a simple, static collection of data. It is large, complex, dynamic, and cannot be fully understood without considering the processes involved in its collection, management, operation, and governance.

This section of the report will outline the components of the dataset, the IT environment that supports it, the major functions involved in processing and managing NDS data, and the issues and challenges that would have to be addressed in migrating this data set to another custodian.

5.1. Components of the Dataset

There are three major components of the NDS dataset:

- A database that contains information about various entities involved in the study, e.g., drivers, vehicles, trips, events; and the relationships among these entities.
- Video files captured by in-vehicle cameras, which are linked to entities in the database.
- An archive of raw sensor data as captured by in-vehicle data acquisition systems (DAS).

Each of these components is discussed in turn below.

5.1.1. Database

The database is largest component of the NDS dataset, consuming approximately 1.5 petabytes (PB) of logical storage space. (Actual physical space utilized is greater by a factor of approximately 25% owing to data protection mechanisms that replicate portions of data across multiple disk drives to protect against data loss due to individual drive failures.) To put this into perspective, a typical desktop computer today is equipped with a 1 terabyte (TB) disk drive. 1 PB equals 1,000 TB, so factoring in the additional storage capacity needed for data protection, the NDS database requires the equivalent storage of 1,250 desktop computers.

Storing and managing sensor data alone is a significant challenge, since there are approximately 18 trillion such rows of data (records) spanning 11 tables in the database.

The database runs on IBM's Db2®relational database management system (RDBMS), considered an enterprise-class RDBMS because of its feature set and scalability. In the NDS environment, Db2 software runs on a cluster of 5 high-performance servers (4 primary, with 1 standby for failover recovery). Using a cluster of database servers is necessary for a database of this scale since both data and processing workload can be distributed transparently across multiple nodes, e.g., splitting a complex query into component tasks that execute in parallel.

5.1.2. Video Files

Internal and external facing video captured from the study vehicles are stored in a high-performance, network-attached filesystem and consume approximately 1 PB of logical storage space. The video files themselves are organized in a hierarchical series of directories based on collection site, source and timestamp. Individual files are linked to logical entities in the database, such as trips and events, based on references to files that are stored in the database.

5.1.3. Archival Storage

The original raw data as captured by in-vehicle DAS units was periodically offloaded and loaded into archival storage before being ingested into the database. These data consume approximately 1 PB of storage capacity and need to be considered a logical part of the NDS dataset for several reasons, including providing traceability to original sources, and offering a potential means for recovering a significant portion of the database rapidly in the event of a catastrophic failure.

5.2. Major Processing and Management Functions

The NDS dataset can be thought of as a collection of content that requires on-going curation. Some of the activities involved in the curation of the NDS dataset are described below.

5.2.1. Data Reduction

Data reduction is the transformation process that derives additional information and meaning from digital data and video in the NDS dataset. Events such as crashes, near crashes, baseline driving, and myriad other categories have been cataloged by algorithmic or manual processes, or a combination of both, and event cataloging continues to this day. Additionally, annotations are created and associated with events. Events and annotations accumulate as research involving the NDS dataset continues, thus the database continues to grow.

A proprietary VTTI application known as “Hawkeye” (and related middleware) is used by VTTI personnel, as well as researchers whose DUL allows them to work in an SDE, to query the database and view video that is time-synchronized with the event(s) being analyzed. This application also provides a function to add annotations to events.

5.2.2. Dataset Extraction

MATLAB® is the principal software tool used by VTTI analysts to identify and extract data based on requirements worked out with researchers, usually through an iterative process. Over time, a significant set of MATLAB scripts have accumulated at VTTI, representing a valuable collection of reusable intellectual property.

An important aspect of the IT infrastructure associated with the NDS data set is the high-performance compute cluster and 10-gigabit network backbone at VTTI. This provides parallelization of MATLAB jobs for dataset extraction, and efficient execution of intensive machine learning and image processing algorithms against the entire dataset.

5.2.3. Collection Administration

Collection Administration is an easy to overlook but essential facet of NDS dataset curation. VTTI have an internal system that is used to manage information about driving studies. Each driving study in which VTTI has been involved represents a “Collection” in this internal, administrative database. The SHRP2 Naturalistic Driving Study is such a collection. The data managed related to a study, or collection, include detailed driver and vehicle information, and driver survey responses.

Appropriate, non-PII data in the collection administration database have been replicated in the NDS Db2 database. Highly sensitive PII must remain segregated for governance and legal reasons.

This assessment effort uncovered the fact that researchers occasionally request data elements that can only be found in the collection administration database, not because they contain PII, but simply because they were not thought to be significant for research purposes. One such example is vehicle width, which turned out to be an important vehicle attribute for a lane keeping study. Any data migration effort would therefore have to include a task to identify potentially valuable non-PII data that should be replicated into the NDS dataset.

5.2.4. Insight

Insight is a website available to registered users at https://insight.shrp2nds.us that provides, among other resources, a repository of de-identified data and video files that is updated periodically from the NDS dataset. These data, along with background and descriptive information available on the site, have enabled researchers to formulate research plans, dataset extract specifications, and even conduct studies. Any movement of the NDS dataset to another custodian must consider the value of Insight to the research community.

5.3. Challenges

Numerous challenges would have to be overcome to migrate the NDS dataset. The most significant ones are outlined below.

5.3.1. Hardware and Software Acquisition and Maintenance Costs

As noted previously, a significant amount of high-performance compute, storage, and networking hardware are required to support the NDS dataset. Hardware acquisition costs alone would be in at least the 7-figure range, with significant recurring costs for maintenance services, facilities, power, cooling, and systems management staffing. Operating budgets would also have to account for the 3 to 4-year useful life of most hardware components.

Software licensing costs would also be significant. Db2 licensing costs for a configuration similar to the current environment could approach $1M, with an additional 18-20% cost per year for support and maintenance.

> A Note About Cloud Services. Moving the NDS dataset to “The Cloud” was raised as a possibility for its long-term sustainability given the attractiveness of its pay-as-you-go model. Cloud services are most cost-effective for applications that combine highly elastic compute requirements with either infrequently accessed storage or limited storage needs. Large amounts of high-performance storage are very costly in a cloud environment, and managing the NDS dataset requires dedicated compute, network, and storage resources regardless of where the data “lives”. Absent a cloud computing company providing resources on a free or highly discounted basis, this has not proven to be viable alternative. Moreover, all of the challenges outlined below would still need to be addressed.

5.3.2. Data Movement

It’s impractical to move 3.5 PB of data over a wide area network. Physical movement of data would have to be accomplished by installing a new storage subsystem(s) in the VTTI data center, copying data to it via a high-speed local network connection, then shipping the storage subsystem to its final destination.

5.3.3. *Database Complexity, Scale, and Time*

The Db2 database provides an essential framework for managing and accessing the NDS data, and its structure and organization must be maintained if the data is migrated.

A “back of the envelope” estimate of the elapsed time to export the database itself is upwards of a year, and it could take another 9 months to a year to import the data into a new database environment.

There remains the thorny problem of the dataset changing as new events and annotations are added, and linkages are made to other related information, such as weather data.

5.3.4. *Professional Services Costs*

Any data migration effort would require significant person-hours to accomplish. Roles and skill sets involved would include project management, plus systems, network and database administration. Support from hardware and software vendors might be required. Like other costs, these can only be estimated after an exhaustive, bottoms up evaluation, but a high 6-figure amount is a conservative (low) estimate.

5.3.5. *Intellectual Property*

A substantial set of software tools (applications, middleware, and scripts) are part of the overall SHRP2 NDS data “ecosystem”. The intellectual property rights to many of these tools belong to VTTI because they were originally developed for driving studies that pre-date the SHRP 2 program. The NDS data are of limited use without these tools or their equivalent.

6. Conclusions

Migrating the SHRP2 NDS dataset is not a simple task of copying data from point A to B. The data span multiple platforms, and the data, tools, and processes involved are inextricably linked. The effort and costs involved are considerable, and would constitute a multi-year, multi-million-dollar undertaking.

To be of on-going value to researchers any new environment must provide functional equivalence to the current one and be able to do so without disruption.

While it might be technically feasible to move the NDS dataset, considering all of the complexities, costs, time, potential disruption to users, and unknowns discussed herein, it is not economically feasible to consider this a viable option given the ongoing use of the data by qualified researchers and the budgetary resources available to support them at this time.