International Journal of Population Data Science

Journal Website: www.ijpds.org

Administrative Data Format Standardization for Efficient Analytics

White, RM[1*]

[1]Statistics Canada

Adoption of non-traditional data sources to augment or replace traditional survey vehicles can reduce respondent burden, provide more timely information for policy makers, and gain insights into the society that may otherwise be hidden or missed through traditional survey vehicles. The use of non-traditional data sources imposes several technological challenges due to the volume, velocity and quality of the data. The lack of applied industry-standard data format is a limiting factor which affects the reception, processing and analysis of these data sources. The adoption of a standardized, cross-language, in-memory data format that is organized for efficient analytic operations on modern hardware as a system of record for all administrative data sources has several implications:

- Enables the efficient use of computational resources related to I/O, processing and storage.
- Improves data sharing, management and governance capabilities.
- Increases analyst accessibility to tools, technologies and methods.

Statistics Canada developed a framework for selecting computing architecture models for efficient data processing based on benchmark data pipelines representative of common administrative data processes. The data pipelines demonstrate the benefits of a standardized data format for data management, and the efficient use of computational resources. The data pipelines define the preprocessing requirements, data ingestion, data conversion, and metadata modeling, for integration into a common computing architecture. The integration of a standardized data format into a distributed data processing framework based on container technologies is discussed as a general technique to process large volumes of administrative data.

*Corresponding Author:
Email Address: ryan.white4@canada.ca (RM White)

http://dx.doi.org/10.23889/ijpds.v3i5.1045