A Nonrelational Data Warehouse for the Analysis of Field and Laboratory Data From Multiple Heterogeneous Photovoltaic Test Sites

Yang Hu, *Member, IEEE*, Venkat Yashwanth Gunapati, Pei Zhao, Devin Gordon, Nicholas R. Wheeler, Mohammad A. Hossain, *Member, IEEE*, Timothy J. Peshek, *Member, IEEE*, Laura S. Bruckman, Guo-Qiang Zhang, *Member, IEEE*, and Roger H. French, *Member, IEEE*

***Abstract*—A nonrelational, distributed computing, data warehouse, and analytics environment (Energy-CRADLE) was developed for the analysis of field and laboratory data from multiple heterogeneous photovoltaic (PV) test sites. This data informatics and analytics infrastructure was designed to process diverse formats of PV performance data and climatic telemetry time-series data collected from a PV outdoor test network, i.e., the Solar Durability and Lifetime Extension global SunFarm network, as well as point-in-time laboratory spectral and image measurements of PV material samples. Using Hadoop/HBase for the distributed data warehouse, Energy-CRADLE does not have a predefined data table schema, which enables ingestion of data in diverse and changing formats. For easy data ingestion and data retrieval, Energy-CRADLE utilizes Hadoop streaming to enable Python MapReduce and provides a graphical user interface, i.e., py-CRADLE. By developing the Hadoop distributed computing platform and the HBase NoSQL database schema for solar energy, Energy-CRADLE exemplifies an integrated, scalable, secure, and user-friendly data informatics and analytics system for PV researchers. An example of Energy-CRADLE enabled scalable, data-driven, analytics is presented, where machine learning is used for anomaly detection across 2.2 million real-world current-voltage *(I–V)* curves of PV modules in three distinct Köppen–Geiger climatic zones.**

***Index Terms*—Data science, degradation science, Hadoop, HBase, informatics, photovoltaics (PVs).**

I. Introduction and Background

MOST utility-scaled photovoltaic (PV) systems are instrumented with meteorological and irradiance sensors to monitor the system performance and weather conditions according to PV system monitoring guidelines [1], [2]. These information-rich, temporally continuous or semicontinuous datasets are often complemented by high-resolution solar data and other climate data from nearby regional weather stations measurements, and TMY3 data in the National Solar Radiation Database [3]. In addition, irradiance data can be retrieved from geographic information system (GIS) satellite images; an example is the commercial datasets from SolarGIS [4]. PV system monitoring has proven useful over many years, and more specific methods for PV energy data analysis are needed to improve the performance of new installations and improve the management of existing power plants [5]. Recent durability studies of PV materials show that a more scalable data warehouse and analytics infrastructure is enabling vast new insights [6]–[8], and the new approach to degradation science of power systems raises the demand for large-volume high-temporal-resolution real-world data [6].

At the Solar Durability and Lifetime Extension (SDLE) Research Center, a "Global SunFarm Network" (GSFN) was established with 16 outdoor test facilities in six countries [9]. As shown in Table I, these sites are heterogeneous in terms of plant topologies and hardware. In order to process, store, and analyze about 120 GB of time-series data that accumulate from the GSFN each year and integrate them with laboratory-based data, we need a data informatics infrastructure that is dedicated to solar energy research. Existing solar energy databases are based conventionally on relational database management systems (RDBMS), for example, several databases supported by the National Renewable Energy Laboratory (NREL) in the Open PV Project [10], and PV Data Acquisition project [11], as well as the International Energy Agency's Photovoltaic Power Systems Program performance database [12]. The initial data infrastructure design of the SDLE GSFN originates from NREL's proposed regional test center (RTC) data warehouse proposal [13], which uses a typical RDBMS MySQL database.

Several drawbacks of the initial design influenced the performance and efficiency of the system. First, MySQL databases require a static database schema, which limits the scalability of the database structure and increases the preprocessing overhead of ingestion. Moreover, the proposed RTC schema required having individual tables for each data type, which is costly to construct and to maintain. As GSFN expanded to more test facilities and commercial PV power plants, the poor scalability of the RDBMS data model and accumulation of a large number of tables would be an obstacle in the near future. Furthermore, sequential data processing would be delayed resulting in longer database write times, due to the atomicity, consistency, isolation, and durability properties of RDBMS databases [14].

Manuscript received July 18, 2016; revised September 29, 2016; accepted October 31, 2016. The SDLE Research Center was established through funding through the Ohio Third Frontier, Wright Project Program Award Tech 12-004. The Energy-CRADLE project was supported by the Bay Area Photovoltaic Consortium Prime Award DE-EE0004946 under Subaward Agreement 60220829-51077-T.

The authors are with the Solar Durability and Lifetime Extension Research Center, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: yang.hu@case.edu; vxg120@case.edu; pxz83@case.edu; dag109@case.edu; nrw16@case.edu; mohammad.a.hossain@case.edu; tjp3@case.edu; lsh41@case.edu; gqatcase@gmail.com; roger.french@case.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JPHOTOV.2016.2626919