Integrated Large-Scale Data Management Platform for Photovoltaic Power Conversion Equipment (PCE) Reliability Data

Liwei Wang[1], Buck Brown[1], Shuan Dong[5], Tan Jin[5], Daniel Clemens[2], Joseph Hodges[3], Adam Reeves[4], Josh Ozbeytemur[4], Shuangshuang Jin[1], and Zheyu Zhang[1]

[1]Clemson University, North Charleston, SC, 29405, USA

[2]SMA Solar Technology AG, Niestetal, Kassel, 34266, Germany

[3]Dominion Energy, Cayce, SC, 29033, USA

[4]Hannah Solar Government Services, Summerville, SC, 29483, USA

[5]National Renewable Energy Laboratory, Golden, CO, 80401, USA

2023 IEEE 50th Photovoltaic Specialists Conference (PVSC) | 978-1-6654-6059-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/PVSC48320.2023.10359922

***Abstract* — To meet the demand for accuracy and real-time capability of PV system degradation evaluation, massive volume data is needed to run high-fidelity and high-efficiency simulations and perform advanced data analysis. However, PV farm operators have a series of difficulties with PV inverter data, such as data collection from multiple channels, massive data storage, data management and massive data analysis. To address these challenges, we developed an integrated data management platform capable of data acquisition, processing, storage, query, and performing big data analysis utilizing AI algorithms. The platform can also achieve data correctness verification and provide an effective distributed data management solution to retrieve massive data and establish a connection to distributed computational frameworks.**

Keywords— data management platform, distributed computing, field reliability data, PV inverter

I. Introduction

The installation of PV systems has grown across the globe in the first half of 2022 [1]. Influenced by the enactment of the Inflation Reduction Act (IRA), the increase of photovoltaic (PV) penetration into the U.S. power market can be estimated to be further incentivized in the future. As more PV systems come into service rapidly, component operation and maintenance (O&M) costs need to be noticed. It is evidenced by field data from PV power plant operators that power electronic converters contribute most to O&M events, responsible for between 43% and 70% of the service calls [2][4]. Field reliability data is an important indicator to help operators to monitor the operational status under measured environmental stressors. Based on the historical performance data, operators can better evaluate the current state of health condition and predict the lifetime of components. Meaningful field reliability data may include parameters of the local environment (temperature, humidity, irradiance), PV farm (configuration considering grounding), grid data (grid command, power quality, grid disturbance), and inverter data (DC-link voltage, P/Q reference). The collection and analysis of field reliability data may help operators to better understand the degradation pattern of inverters. In addition, operators can schedule necessary maintenance in advance, reducing the likelihood of PV system failure.

The traditional field data collecting and managing approaches meet several challenges:

1. Some field reliability data is measured by the built-in measurement tools inside the component. It is only accessible from the designated data portal provided by the manufacturers. Environmental data and inverter operational data are usually collected by distinct channels. There are still several difficulties with integrating data from different sources.
2. The correctness of collected field data is not always dependable. For example, if the solar irradiance sensors are blocked by other objects, the irradiance recorded at that time is inaccurate. It is difficult to verify the correctness of collected field data.
3. More advanced measurement tools with higher accuracy and higher time resolution are applied widely. Massive data with higher precision is accumulated as time grows. The way to efficiently store, retrieve, and analyze mass historical data is also a challenge.

In this paper, we propose an integrated field data management platform that can address the challenges mentioned above. Field data collection is introduced in Section II. The data management platform capability is introduced in Section III. Section IV discusses conclusions and future work.

II. FIELD RELIABILITY DATA COLLECTION

The importance of renewable energy cannot be overstated as the world transitions to net zero carbon emissions. PVs are a type of renewable energy source that have vast potential yet currently face reliability issues. In particular, PV inverters are one of the most unreliable subsystems within the larger PV system due to their complexity. Some of the components within PV inverters that have the worst reliability are capacitors, cooling fans, metal oxide varistors, printed circuit

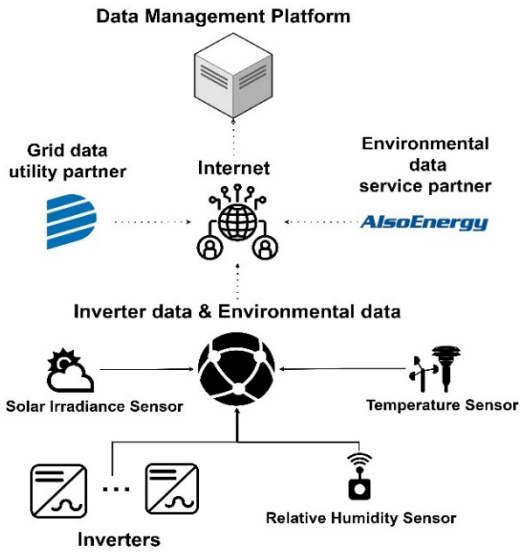

Fig. 1. Field reliability data acquisition dataflow.

boards, power modules, and relays/contactors. Collecting data on the stressors that cause these components to fail, and thus, the entire PV system to fail, could provide invaluable insight. As such, a field reliability data collection technique that captures stressors such as temperature, humidity, power, voltage, current, etc. is paramount.

For the purpose of understanding the comprehensive performance of PV inverters, a field data acquisition system is established. The structure of the data acquisition system is shown in Fig. 1. Two categories of field data are collected: operating condition data and environmental data. Operating condition data is comprised of grid data and inverter data. Environmental data provides historical environmental conditions around PV inverters.

Grid data is provided by the grid data source. It can be acquired via the online portal provided by the grid data utility partner. The grid data includes three-phase current, voltage, and power. It also includes transient data records to indicate the operating condition data when faults or dynamic events happen on the grid. **Inverter data** includes AC and DC power, voltage, and current. It is recorded by the measurement tool designed in the PV inverters. The data is transmitted to the user interface that can help operators to conduct energy monitoring, managing, and grid-compliant power control. **Environmental data** includes the ambient temperature, relative humidity in the PV farm, Global Horizontal Irradiance (GHI) received by the PV farm, and module temperature around critical electrical components in the PV inverter. These types of environmental data have two data resources. One is from the sensors deployed around the PV inverters. This data is captured, uploaded to the web portal, and known as environmental data source A. The other is from the sensors deployed around the PV farm. This data is captured, uploaded to the web portal, and known as environmental data source B.

To demonstrate the validation of our design, the following case study is conducted. **SMA SUNNY-TRIPOWER-30000TL-US** inverters (30000TL-US-10) are used at the Otarre solar farm in Cayce, South Carolina, USA. Data sources provided by our utility and service partner, their corresponding web portal, and their time resolution can be found in TABLE I.

TABLE I. DATA SOURCE REFERENCE

Data Source	Utility & Service Partner (Online Portal)	Time Resolution
Grid Data Source	Dominion Energy	10 minutes
Inverter Data Source	SMA (Sunny Portal Powered by ennexOS)	5 minutes
Environmental Data Source A	SMA (Sunny Portal Powered by ennexOS)	1 minute
Environmental Data Source B	Also Energy (PowerTrack)	5 minutes

III. Data Management Platform

To overcome the challenge of managing the field reliability data efficiently, we established a data management platform to integrate the data and perform the preliminary analysis based on the collected historical data. It 1) integrates field data from different channels, 2) applies big data solutions to optimize data retrieval and analysis capability, and 3) performs cross-data validation to increase the accuracy and reliability of collected field data. Related technical details of these features will be introduced in the rest of this section.

A. Data Integrations

In our data acquisition system, there are online data portals for grid data, environmental data, and inverter data. Each online data portal provides its own user interface to help operators monitor the operating condition and environmental information as well as track historical data. However, monitoring and tracking historical data across different channels is not supported by any existing online portals. As a result, we developed a third-party, python-based data management system to integrate multi-channel data and provide an efficient data retrieval solution to monitor and track the inverter operating condition and surrounding environmental information.

The workflow of the data management is shown in Fig. 2. At first, dedicated web scraping scripts extract field data from the three online portals separately. The web scrapings simulate the web browser to send requests for target data and gather the target data by fetching the response returned from the web portal server. Request, BeautifulSoup, and Selenium libraries are utilized in this application to perform the data scrap and parse the data.

If some data source is missing, this data is marked with an error flag to be filled with the proper value later. Then, data is inserted into the corresponding MySQL databases and awaits further processing.

B. Data Correlation and Fusion

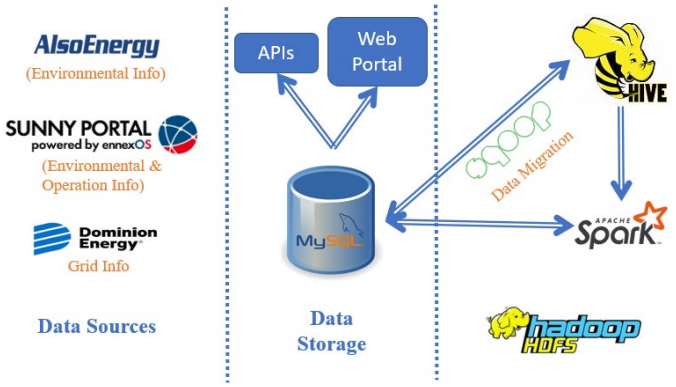

Fig. 2. Data management platform workflow.

The same data field may have multiple data sources, therefore, the measurements from multiple channels can be merged into a more accurate dataset. The list of data that have multiple channels is shown in TABLE II.

Data correlation: To avoid the failure of data fusion caused by invalid source data, several data validation actions are taken to verify the consistency and reasonability of the data. The Module Temperature Check [6] is performed to reduce the measurement noise. Besides, the isolation Forest algorithm [5] is performed to detect outliers. Isolation Forest constructs iTree based on the features of the dataset. The node near the root node of iTree is detected as an outlier. The outlier data is filled with the reasonable value derived based on the nearest data point.

Data fusion: After assuring the validation of the data source, data fusion of multiple data channels can be performed. However, due to discrepancies in the location of sensors, the measurement scope, and the measurement accuracy, the same type of datasets from different channels are not always identical. The field data from Sunny Portal is collected by the dedicated sensors deployed in the PV inverter closure to evaluate the operational status of PV inverters and the environmental conditions around inverters, and hence, the data from Sunny Portal is more reliable. Other data channels are regarded as complementary sources to increase the fidelity and accuracy of the data. As for operational information, like AC power, the power data from Sunny Portal provides the operating condition of each inverter, and the power data from the grid data source provides the operating condition of the grid. When dynamic events are detected on the grid, the data collected by individual inverters is not as accurate as the data from the grid because of higher resolution measurement instrumentation equipped at the grid side by the utility company.

The metadata from different channels is organized in different formats and time resolutions. Therefore, the data from the channel with a lower time resolution can augment to a higher resolution. For example, the temperature from environmental data source A is updated every minute, and the temperature from environmental data source B is updated every five minutes. Temperature data from Sunny Portal is compensated by referring to the data from the other channel.

TABLE II. DATA FIELD AND SOURCE

Information Type	Data Field	Data Source
Environmental	Ambient Temperature (°F)	Environmental Data Sources A and B
	Solar Irradiance (W/m^2)	Environmental Data Sources A and B
Operational	AC Power (W)	Inverter and Grid Data Sources
	AC Voltage (V)	Inverter and Grid Data Sources
	AC Current (A)	Inverter and Grid Data Sources

Compared to the traditional interpolation method, the data from the other channel provides a more accurate estimation. After finishing the data correlation and data fusion, the merged dataset is integrated into the database.

C. Data Management

With the field data collected over time, proper data management solutions are needed. To satisfy the demand for data storage, query, and further analysis, the relational database, distributed data framework, Apache Hadoop, and other techniques are utilized. These utilization details are discussed in the rest of this section.

Metadata management: Due to the need for the prompt query for a large number of metadata, the field metadata is stored in the relational database MySQL. The field data can be updated and retrieved via SQL commands efficiently. For the convenience of queries, an online demo portal is developed to acquire specific data as shown in

Fig. 3. APIs are also provided for querying the field data from the database with the specific data field in the designated time range. The metadata can be easily acquired for future R&D needs.

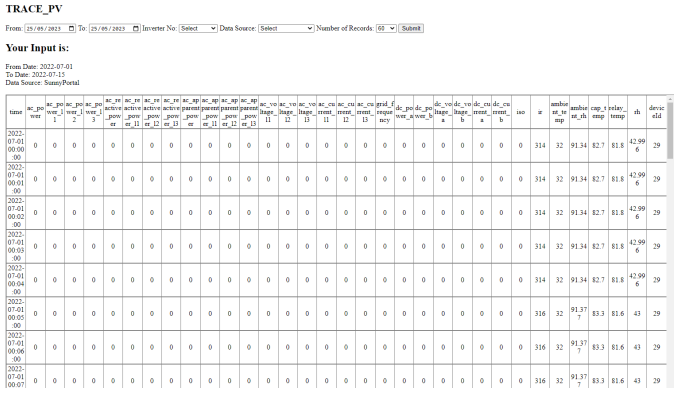

Fig. 3. Metadata query online portal.

Distributed data management: As the metadata accumulates, performing data processing and analysis on massive data becomes more expensive. It takes a long time to read data, perform computation, and write data back to the database when the data volume becomes large. Therefore, distributed data management solutions are needed. The Hadoop Distributed File System (HDFS) is a distributed data

storage system used in the Hadoop ecosystem [9]. HDFS stores the data across multiple machines or nodes in a cluster. It is the foundation of other Apache applications. Hadoop applications follow the master-slave architecture to achieve storing and processing of massive amounts of data effectively. A portion of data is assigned to each node and the data on each node is replicated across the node for failure recovery. Considering the data update and query efficiency, the metadata is stored in MySQL, but for the convenience of performing distributed data processing and analysis, metadata needs to be migrated to HDFS. Apache Sqoop is used as a translator to migrate the metadata from structured data in MySQL to unstructured data in Apache Hive called Apache Spark. Apache Spark is designed for fast computation in performing data science and machine learning on single machines or clusters. This workflow can be deployed on cloud service platforms, like AWS and AZURE.

IV. Case Study

A case study is performed to demonstrate the efficiency and feasibility of the workflow we designed. We selected the operating condition data for the same inverter with a time resolution of 5 minutes and environmental condition data with a time resolution of 1 minute as a use case to introduce the workflow of the data management platform. All the data is measured at Ottare PV farm located in Cayce, South Carolina. To be specific, we select the data for a whole day on May 15th, 2022, to introduce the data correlation and fusion workflow. And we select the data for a whole year from May 2022 May 2023 to demonstrate the distributed data management.

Both operating condition data and environmental data are collected by web scraper scripts from three data resources daily. The collected data will do the data correlation. Module Temperature Check will be used to evaluate the validation of the temperature and irradiance information. The completeness of data will be evaluated. Empty, missing data and outlier values will be filled with the value measured at the closest moment if the completeness is over 99% (The ratio of Nan value to total values). Or else, also including other situations for data fusion, it will be reported as the missing data period. The data from the second data channel will be used and filled into the missing data period. Because the data from two different channels vary caused of location differences, the similarity between the data from the two channels will be evaluated. If the cosine similarity between data from two channels is lower than 99.5%, the second channel data will be converted based on the gap between the data from the two channels before being filled. And the data from the second channel will be filled directly if the data from the two channels are similar. The comparison of ambient temperature data before and after data correlation and fusion is shown in Fig. 4. After the data correction and fusion, metadata will be stored and managed by the relational database MySQL for prompt data queries.

When the data accumulates to a large volume, the data manipulation and advanced data analysis are very time-consuming. In order to increase the operation efficiency, the metadata will be migrated from MySQL via the Sqoop to HIVE for later use by Apache SPARK. Advanced statistics can be performed via PySpark more efficiently compared to traditional data management platforms.

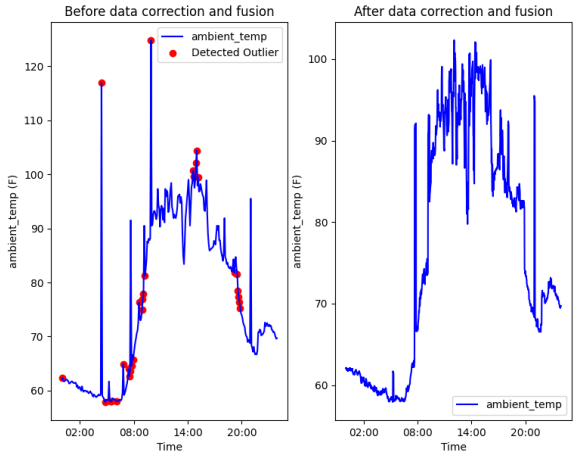

Fig. 4. Comparison of results before and after data correction and fusion.

Besides, we also can perform machine learning algorithms using Mllib [8] for fast advanced analysis when handling large-scale data. For example, cluster algorithms can be used to categorize environmental condition data of different days into different weather patterns. To better visualize the clustering result, the data has been decomposed into two-dimensional data, which is shown in Fig. 5. The red points are the locations of the cluster centers. The blue points are locations of the one-day environmental conditions mission profile in two-dimension. The clustering result can help to merge similar mission profiles and decrease the simulation workload by skipping using similar mission profiles as input.

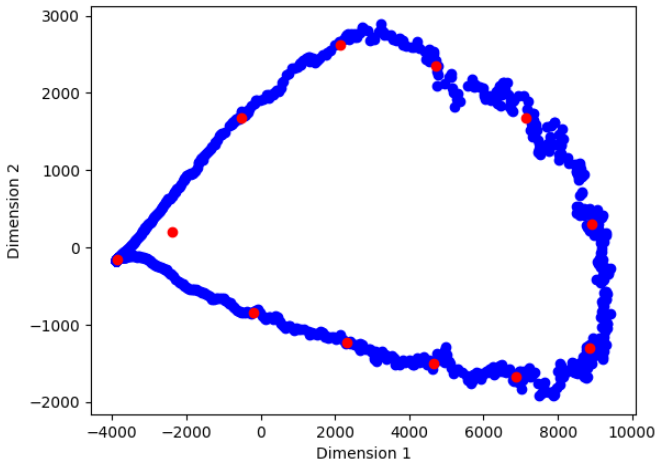

Fig. 5. Clustering results for environmental conditions of one year.

V. Conclusion and Future Work

To meet the demand for accuracy and real-time capability of PV system degradation evaluation [10], massive volume data is needed to run high-fidelity and high-efficiency simulations [11] and perform advanced data analysis. In this paper, the implementation details of field data management are introduced. And a demo example use case is described to show the feasibility. **our main contribution is developing an integrated data management platform capable of data acquisition, processing, storage, query, and performing big data analysis utilizing AI algorithms.** The platform can also achieve data correctness and provide an effective distributed data management solution to retrieve massive data and establish a connection to distributed computational frameworks. This data management platform provides a large-scale computation capability for PV-related R&D needs. Existing analysis tools and algorithms which are subject to computation capability and large-scale data manipulation efficiency can also be moved to this platform. For the current stage, we focus more on the environmental condition data process and analysis. In the future, we will add some support targets to the data process and analysis on the operating condition data process.

Acknowledgment

This work is supported by the Solar Energy Technologies Office, Office of Energy Efficiency and Renewable Energy, Department of Energy, USA under the award number DE-EE0009348.

References

[1] V. Ramasamy, J. Zuboy, E. O' Shaughnessy, D. Feldman, J. Desai, M. Woodhouse, P. Basore, R. Margolis, "U.S. Solar Photovoltaic System and Energy Storage Cost Benchmarks, With Minimum Sustainable Price Analysis: Q1 2022", No. NREL/TP-7A40-83586, National Renewable Energy Lab. (NREL), Golden, CO (United States), 2022.

[2] F. David, K. Dummit, J. Zuboy, and R. Margolis, "Spring 2022 Solar Industry Update", National Renewable Energy Lab.(NREL), Golden, CO (United States), 2022.

[3] J. Flicker "Reliability of power conversion systems in photovoltaic applications," Sandia, NW, 2015J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[4] P. Hacke, et al. "A status review of photovoltaic power conversion equipment reliability, safety, and quality assurance protocols." Renewable and Sustainable Energy Reviews 82 (2018)

[5] F. Liu, K. Ting, and Z. Zhou. "Isolation forest." 2008 eighth IEEE international conference on data mining. IEEE, 2008.

[6] Perry, Kirsten, et al. PVAnalytics: A Python Package for Automated Processing of Solar Time Series Data. No. NREL/PR-5K00-83824. National Renewable Energy Lab.(NREL), Golden, CO (United States), 2022.

[7] Aravinth, S. S., et al. "An efficient HADOOP frameworks SQOOP and ambari for big data processing." International Journal for Innovative Research in Science and Technology 1.10 (2015): 252-255.

[8] Meng, Xiangrui, et al. "Mllib: Machine learning in apache spark." The Journal of Machine Learning Research 17.1 (2016): 1235-1241.

[9] K. Shvachko, H. Kuang, S. Radia and R. Chansler, "The Hadoop Distributed File System," *2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, 2010, pp. 1-10, doi: 10.1109/MSST.2010.5496972.

[10] I. Vernica, H. Wang, F. Blaabjerg, "Design for reliability and robustness tool platform for power electronic systems—Study case on motor drive applications," in 2018 IEEE Applied Power Electronics Conference and Exposition (APEC), 2018.

[11] L. Wang, R. Thiagarajan, S. Jin, and Z. Zhang, "Accelerating Simulation for High-Fidelity PV Inverter System Reliability Assessment with High-Performance Computing". In 2022 IEEE 49th Photovoltaics Specialists Conference (PVSC) (pp. 0178-0182). IEEE.