

Massive Dataset Challenges and Solutions

Alasdair Crawford, Vish Viswanathan, Steve Willard, Joe Thompson, Charlie Vartanian, Dave Reed, Vince Sprenkle

Pacific Northwest National Laboratory, Richland, WA 99352



Pacific Northwest
NATIONAL LABORATORY

Proudly Operated by **Battelle** Since 1965

Introduction: The EPRI/PNNL joint reliability data project's BESS recorded performance data sets require labor intensive pre-processing, before analysis and findings can take place. The project's data management challenges are driven by massive BESS performance data sets, and we outline solutions that are in process and proposed. These solutions include EPRI/SNL's Data Requirements reference document aimed at standardizing reported performance data, and proposed development of massive-data hosting and data management tools aimed at pre-processing and screening data quality enabling more time to be spent extracting useful findings from data.

Objectives:

- Mine data from a variety of energy storage systems to analyze and understand their performance and reliability
- Standardize data storage and collection
- Identify standard analysis techniques on actual field data
- Share these techniques and algorithms with the energy storage community

Approach:

- EPRI has acquired large amounts of data (over 2 years) from energy storage systems (both flow and Li-Ion) and developed a data science platform (DIAMOND) that standardizes its format and storage
- EPRI/Sandia developed data collection document identifying each required data point, what level it is required at (system, rack/module, cell), system specific data points for different technologies, data collection and storage procedures.
- DIAMOND can be queried using SQL based on site, time range, and quantity desired:
 - Power (AC/DC)
 - SOC
 - Temperature
 - SOH
 - Various flags for alarms
 - Bonus information such as ambient temperature, PV plant load
- Data time stamp uses standard ISO 8601 format – completely removes hassle of dealing with time zones, daylight savings, parsing.
- Each of these quantities is typically available at several levels – the system level, the rack/module level, and sometimes the cell level. Furthermore there are typically different summary variables available – such as max, min, and average. This is all in a standardized format, ie for all systems it is essentially the same process to query “What is the average temperature for each rack?”

SystemNode	SystemNodeType	ChannelCatalog	ChannelCatalogUnit	ChannelAggregation	Descriptor
Container 1, Rack 9	Battery Rack	StateOfCharge.SOH	%	(None)	N/A
Container 1, Rack 9	Battery Rack	DC.Voltage	V	(None)	Avg of Cells
Container 1, Rack 9	Battery Rack	DC.Voltage	V	(None)	Min Cell
Container 1, Rack 9	Battery Rack	State.LGChem.ModuleID		(None)	Min Cell V
Container 1, Rack 9	Battery Rack	DC.Voltage	V	(None)	Max Cell
Container 1, Rack 9	Battery Rack	State.LGChem.ModuleID		(None)	Max Cell V


Example of channel organization

- EPRI performed analysis to understand how system was calculating SOC and SOH, while PNNL concentrated on how to predict SOC and SOH – overlapping efforts that fed into each other.

Results and Discussion:

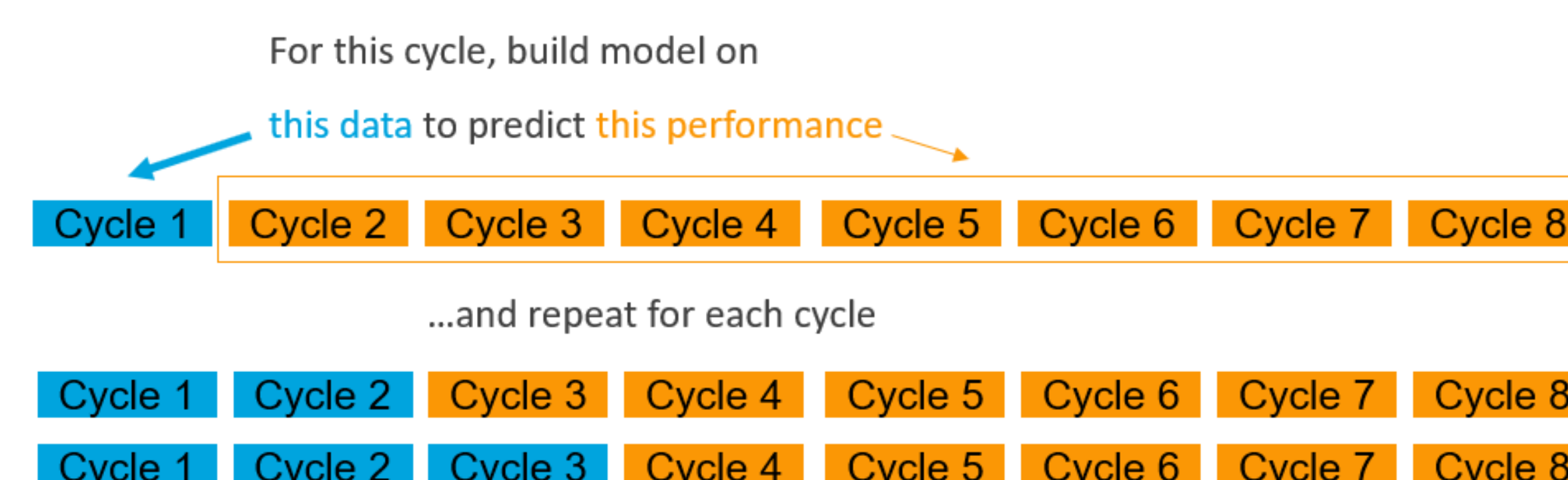
- For purposes of predicting SOC and SOH, a time resolution of ~5 minutes is sufficient. For analyzing internal resistance and response time, time resolution needs to every second – challenging to store and process this much data.
- Data was downloaded from EPRI and stored on PNNL cloud services. Datasets are on the order of 1 GB each with the 5 minute interval.
- Analysis easier if data is transformed from long to wide format

Time	SOC Rack 1	SOC Rack 2
2020-09-15 14:00	20	40
2020-09-15 15:00	30	50



Time	SOC	Rack
2020-09-15 14:00	20	1
2020-09-15 14:00	40	2
2020-09-15 15:00	30	1
2020-09-15 15:00	50	2

- All data is stored in this manner on PNNL side, transformation is done using reshape library in R. Analysis of SOC change as function of time, power, temperature, and SOC straight forward to do for each rack individually or all together in this data format.
- Data is typically cleaned by filtering out everything but charge/discharge cycles where SOC changes by some threshold, typically 10%
- Linear regression and machine learning models used to predict SOC and SOH, and evaluated on their ability to predict SOC and SOH of future cycles:



Future Work:

- Data analysis scripts to be made publicly available on GitHub for Energy Storage community to access
- Graphs summarizing PNNL analysis to be made available in same repository
- Data to be stored in internal SQL database for internal collaboration

Acknowledgements

This work is supported by the U.S. Department of Energy (DOE) Office of Electricity Delivery and Energy Reliability under contract No. 57588. PNNL is operated by Battelle Memorial Institute for the DOE under contract DE-AC05-76RL01830.