In future years, it should become easier to install more sensors in nuclear power plants, providing an enormous amount of data to get deep insight in the process and improve O&M activities for the plant. An efficient strategy to manage the data is to store the data, check its quality and make the right data available for analysis with a broad range of methods. The Halden Reactor Project (HRP) has many years of experience of providing an understanding of data for decision making.

With more data there will be an increasing need for enterprises to organize and structure the data in a practical and efficient way. Traditionally many data sources are governed into silos and made unavailable between those data zones. With new analytic tools in combination with modern computing power and storage capacity it is possible to make operational and experimental data available for analysis. However, most nuclear industry systems still rely on traditional data warehouse structures and Extract-Transform-Load (ETL) for onboarding data. Traditional databases scale poorly as the complexity and the amount of data is increasing, and new approaches for data management, such as data lake frameworks, can meet the need for a flexible storage with better performance scaling. Good strategies to manage the data is necessary to obtain the right information from the data.

The HRP Data Lake project will secure the historical data from the Halden Boiling Water Reactor (HBWR) as well as enable data scientist to perform new types of analysis on the data. The project is an ongoing digitalization, structuring and data quality assurance job on over 50 years of data from the HBWR’s entire lifetime. The amount of data is approximately 1 TB and contains all process and experimental data, existing both on paper copies and in digital form. There are 8 main data source types with more than 60 sub-types in total, including data in several formats, both structured and non-structured such as images, diagrams and reports. The variety of data types makes a standard ETL process non-ideal. Thus, schema-on-load methodology facilitates a more efficient process of loading and working on our data. The project has setup a local cluster with a containerized Big Data and open source data lake solution based on Apache foundation projects. This proof-of-concept setup allowed us to evaluate data lake as a technology and get an operational understanding necessary to set requirements for a production ready solution. Further in the project a validated environment will be acquired and implemented. Data scientists will have several standardized methods of working with the information, e.g. directly using a SQL-like language or collaborative environments such as notebooks with support for common languages, e.g. Python, R, Java, Fortran.