LTI
Let's Solve

Point of view

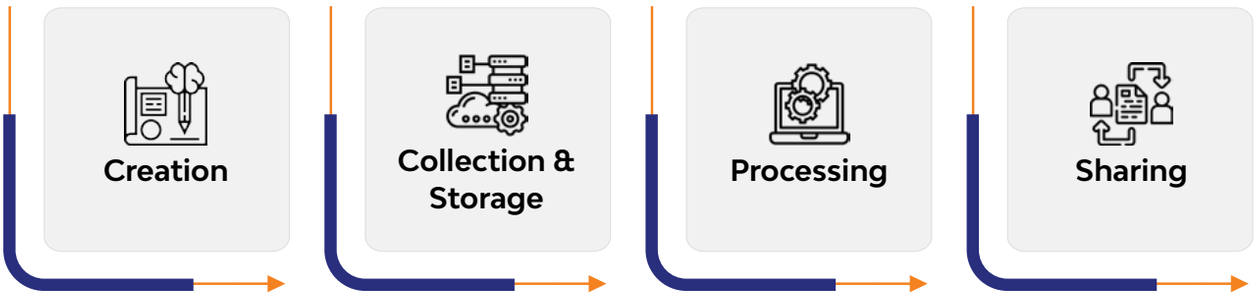
Data Management – Mining and Processing for Superior Business Outcomes

A Larsen & Toubro Group Company

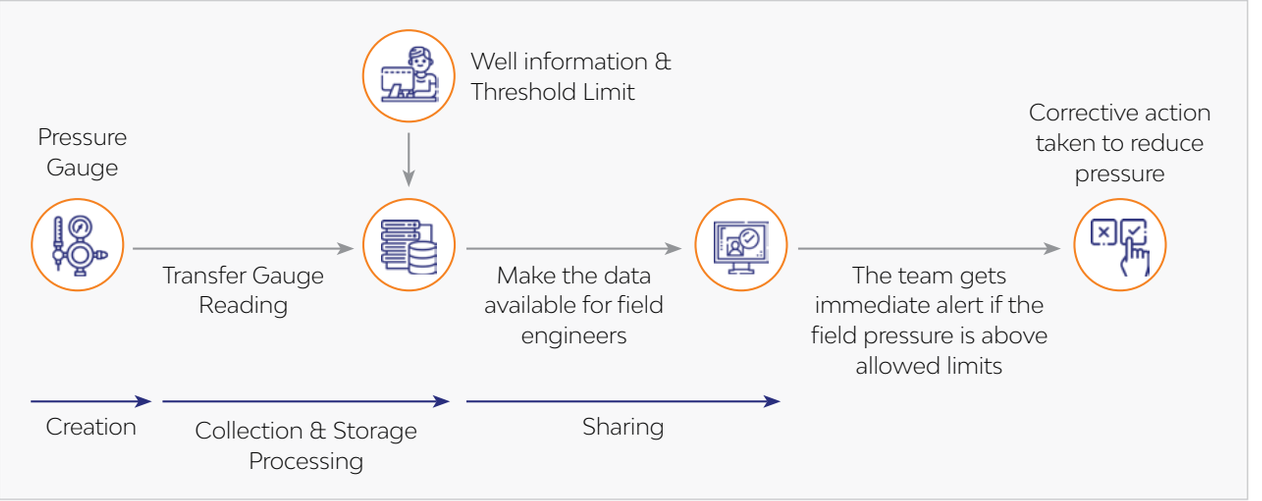
Data is one of the most critical assets of an organization in the current business environment along with its continuously evolving technology; many of an organization's objectives and some of its latest technologies, such as AI (Artificial Intelligence) and ML (Machine Learning), cannot be utilized if they do not possess a significant quantity and quality of historical data.

According to a KPMG CIO survey, only 17 percent of the organisations say they are very effective at maximising the value from the data they hold [1]. A deeper analysis of the organisational systems will also reveal that the volume of data collected is exponentially large and rapidly growing as compared to the growth of IT infrastructure, required to manage the collected data.

Data Management is critical across all business processes within any organization. The process of data management can be broken down into four phases:



Let us consider a simple example of an **onshore field** where oil is extracted using steam flooding. In a steam flood operation, maintaining the correct level of steam pressure in the steam chest is critical. In all cases, the engineers and other critical staff have a threshold limit that needs to be maintained for each field. In addition, there is a pressure gauge at every steam injector, and these gauges generate pressure readings based on the configured settings. In today's data management world, these simple pressure readings prove huge cost-saving with simple data management techniques, as illustrated in the diagram.



The net result of the well-managed, processed, and maintained data is a cost-benefit along with maintaining a safe work environment.

Let us now investigate each of the phases in detail.

Creation: Where is the data created?

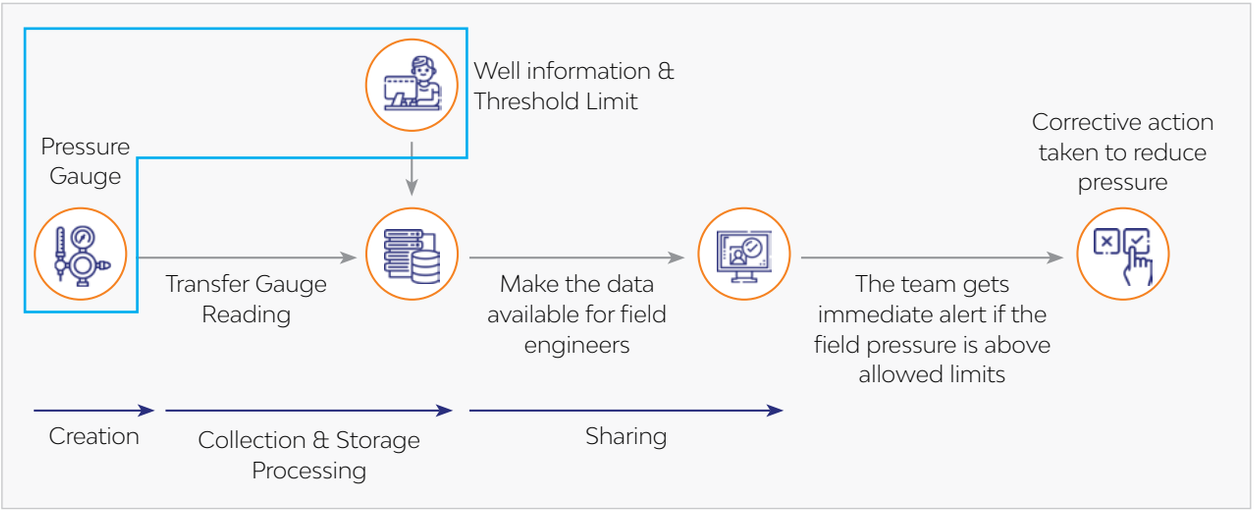
Almost all systems around us generate data. Every device generates some form of data from the instrument installed on a field to the network across which the data flows. From an organizational standpoint, data is created through three main categories:

a. **User entered**
   User entered data are the data sets that are entered by the end-users in an application.

b. **System Generated**
   System-generated data are data sets created inside a system through action or derived data set using some functions. In other words, these data sets are not entered by the users but use the user data as its input.

c. **Purchased data**
   Systems generating data at times can be very expensive to install and maintain. Many organizations thus use third-party data providers to obtain such data sets. For example, a third party vendor can provide North American Wage Costs Forecast data to any retail organization interested in understanding the North American price market.

In our earlier onshore field example, the data can be both user entered and system generated.



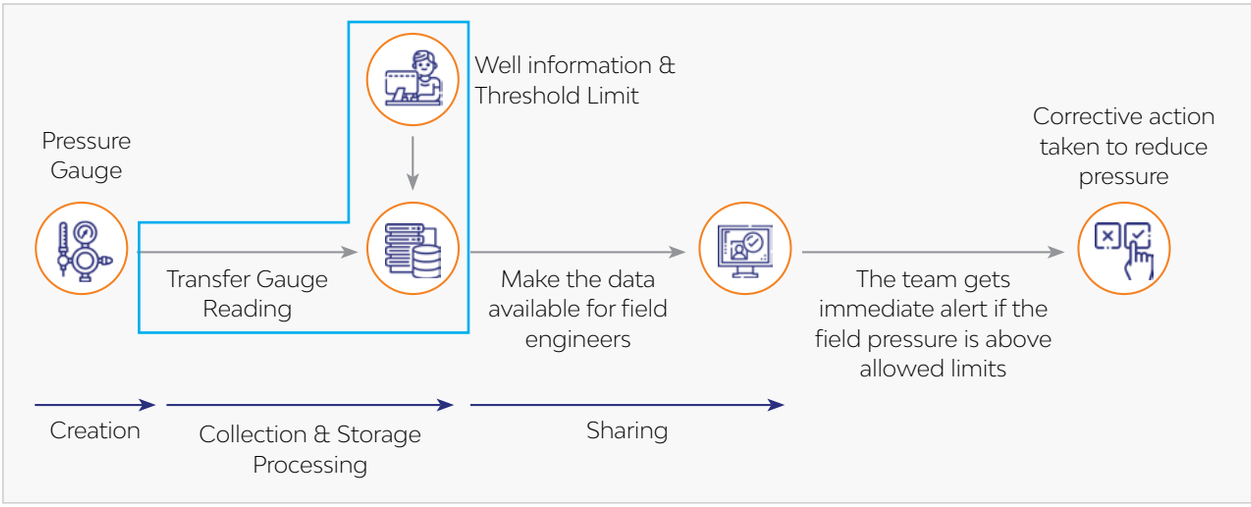
Engineers feed the threshold limit and well information while the gauges automatically generate the current pressure readings.

Collection & Storage: How and where do we store the created data?

Data Creation is only the first phase in the data management process. The second phase is how to collect it and then store it. When it comes to collection points and storage of the created data, the nature of the collected data needs to be understood. For example, a time-based generated data set, aka time-series data, is not manageable when stored in a relational database. Similarly, a relational data set such as low-volume transitional data from a shop floor will be best managed in a relational database. (These are just individual examples and not a recommendation). The second aspect we need to understand is how easy it is to access the stored data by the consumers. There is no point in collecting and storing data in inaccessible storage.

When it comes to purchased data, organizations should also consider the data versioning feature within the storage system for the collected data. Data versioning could be a default feature within the storage system, or it can be built logically. Data versioning allows the organization to restrict the purchased data from a historical trend standpoint. In other words, you need to know what you are paying for.

Let us now look into our example:



In the example, we have two distinct data categories. The user entered threshold limit is a low-frequency transaction. Implying it is not updated on a daily or hourly basis. Therefore, it makes sense to store it in a simple relational database that will hold the field information and the reserve information to which the well is associated. On the contrary, pressure readings are a continuous flow of data values that are transmitted every minute. This would imply that for any given day, we will have 1440 data readings. In a year, we will have 525600 data readings. If we consider the life of the well to be 50 years, then we are talking about 26,280,000 data values. One of the IT systems well suited to manage such data is the OSI PI server. (There can be others also).

From usability and ease of access standpoint, it will be tedious for the user if the threshold and current values are in two different systems. This introduces complexity for a simple business user, technically, it might not sound like a complex case, but this is additional work to integrate the data for an operator. The solution to this would be to hold both kinds of data sets in the OSI PI server.

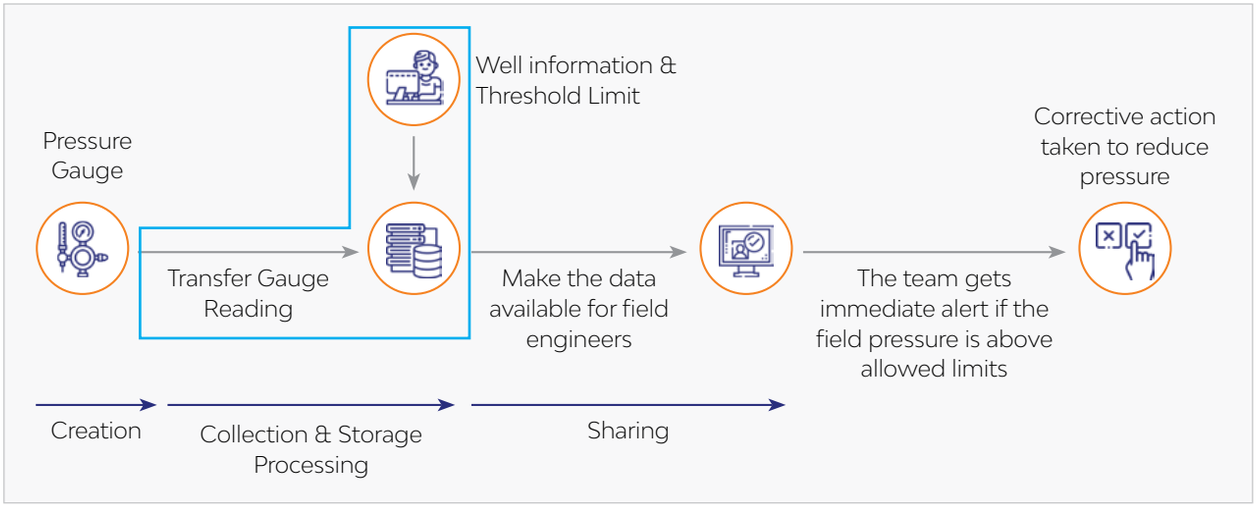
Processing: How do we ensure the data is clean and of good quality?

Once the data is collected and stored, the next stage is processing the data. In an ideal scenario, the data set generated and collected would be the best quality, and often it will not require a quality check. However, this is not the case in reality. Collected and stored data in most cases need a cleansing process. Therefore, it is highly recommended to have a data quality check and correction process, as part of the cleansing phase, right from the very beginning. It is recommended because it becomes more time-consuming and challenging to complete when the amount of data gathered increases.

When considering the data cleansing and quality check process, one can think of a simple categorization strategy. The data quality rules can be used to segregate the data set into three groups. One group will hold the actual raw data, the second group will hold the clean and quality cleared data, and the third group will hold the data set that failed the quality checks. The idea is based on the thought process, "one person's trash is another person's treasure." Many systems use the bad data to generate corrective action items.

In the case of purchased data sets, organizations must build validation steps. These validation steps are needed to ensure the data provided is as per the expected service offering. One of the potential checks is how much data is received and how much is as per the expected quality. In some cases where the pricing is as per data size, a process can be built to calculate data volume.

For example, the data from gauges are system generated and transmitted to the data storage. The ideal condition is that the well operates round the clock without downtime, and the transmission of data is flawless, has zero interference. Unfortunately, in the real world, this is never the case.



Hence data processing logic must be built which can break the collected data set into three sets:

a. Raw data

b. Cleansed data by eliminating zero value readings and conducting other quality checks

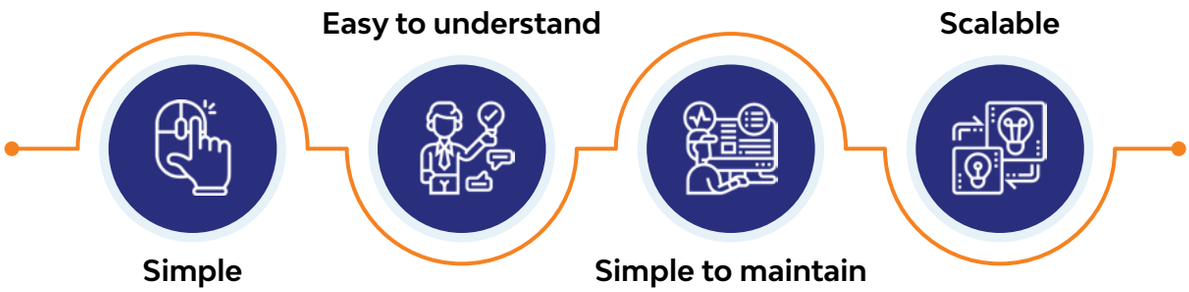
c. Quality test failed data set

The categorization of the data is done without duplicating the records. Instead, it is accomplished by simply adding an attribute to the data set which can hold a flag value.

Sharing: How do consumers get the required data?

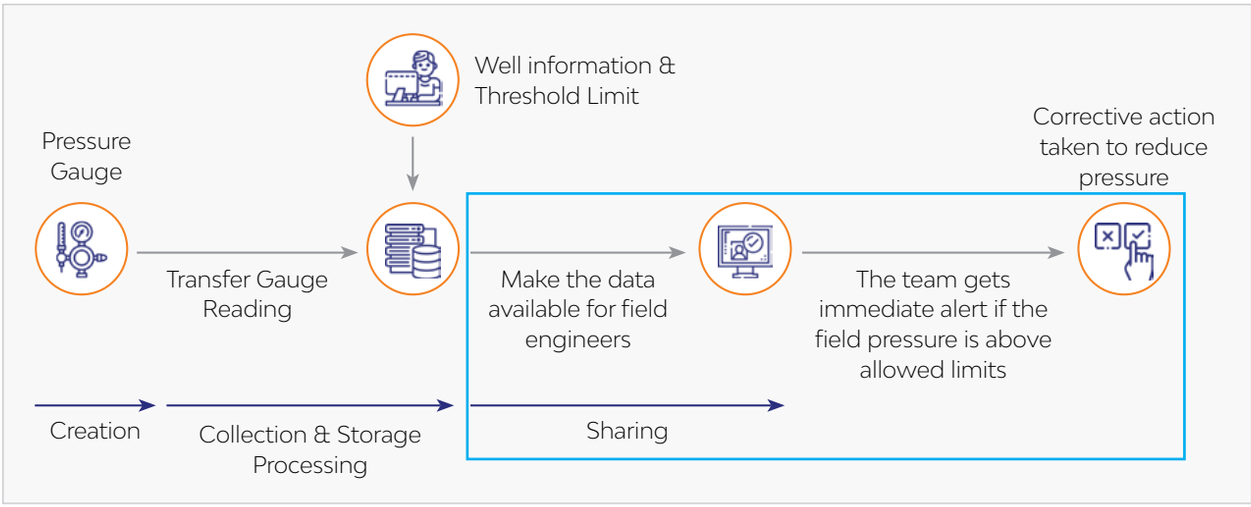
Once the data is created, collected, stored, and processed, the last phase is sharing the data. Organizations can benefit from data only when the data is accessible to the intended user groups. As per a press release by Splunk on September 01, 2020, on "New Industry Research Shows the Volume and Value of Data Increasing Exponentially in the Data Age," it is called out that 66% of IT and business managers believe that more than 50% of their data within the organization is classified under dark data [1]. Dark data is any data set that is untapped, unknown, and typically unused.

Providing access to the stored and processed data can be achieved through a simple step granting users access to the data sets or can be as complex as building a user and application interface for data consumption. To ensure a successful data accessibility interface with a high adoption rate, the interface must be

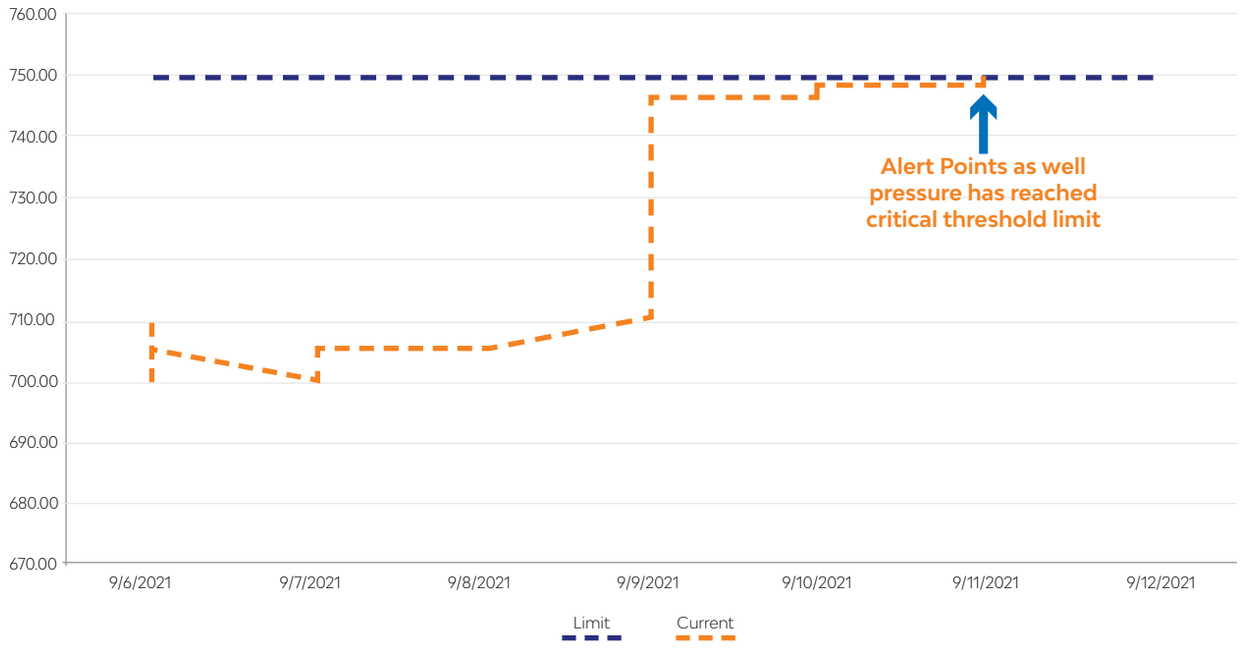


The ease at which an individual can apply analytics and various visualization tools is also a significant measuring parameter for successful data sharing. A good example will be the data set that holds a field's monthly oil production volume. As a field supervisor, one of the most widely anticipated trends is the monthly production volume data. The supervisor can refer to the production trends from multiple meaningful data points for critical decision-making. For example, a declining production volume indicates potential issues in the field or instrument.

With the offshore field example, sharing is when the stored and processed data is made available to the end-users.



When the user gets access to this data set, various analytical software can build numerous types of visualizations to interpret the data. The visualization can help identify or predict anomalies. This helps in acting proactively with preventive maintenance and repairs rather than reactively to the situation, once a breakdown has occurred.



To conclude, in the current information age, the importance and amount of data generated are increasing rapidly. In an overall context, data management is critical as poorly managed data can hinder an organization's ability to make informative and qualitative decisions. Mining existing data and ensuring data quality is also a mammoth task, which can be made easy with new-age technologies of ML and AI.

About the Author

Kevin Vattakatte
Associate Director – Projects, LTI

Kevin is a value-driven and result-oriented versatile professional with more than 15 years of experience in the industry. He is currently responsible for the service delivery to a US-based Oil & Gas client. The role also encompasses leading a digital transformation initiative for the business unit's Upstream value chain business function.

Reference

1. https://home.kpmg/content/dam/kpmg/uk/pdf/2019/11/future-of-it.PDF