Whitepaper

Big data – from data to knowledge
Big data – from data to knowledge

The path of data through the company

1 Introduction

Besides the Internet of Things, handling big data is a basic pillar of Industry 4.0 in the manufacturing industry. Especially in older businesses, there are multitudes of data that cannot be analysed, such as measurement data from the control system or conventional manual inputs. If these data were standardised, stored in a long-term database and built up, real-time analyses and data mining applications could be realised in order to generate knowledge and make predictions. This would minimise downtime and optimise the efficiency of production (even over several links in the production chain).

Based on the current state of production, especially in older businesses, this whitepaper describes the incremental preparation of data and their evaluation – for example, through real-time analyses and data mining applications. Furthermore, it considers other fields of application for data mining, including in areas other than manufacturing.

2 The current state of manufacturing

Although big data, data mining and data analysis have been appearing in the media for years, many companies are still not tapping the new potential. This is primarily due to the long service lives of industrial manufacturing machines before they are modernised or replaced.

As a result, many production facilities still retain old, inhibiting data structures:
The machine control systems write data, yet the data are only stored locally on the PLC. These data only concern local information such as tool changes, type changes, operating times and, of course, measurements. Some of the data are supplemented or modified by manually input information (e.g. at manual workstations).

As the storage capacities of these systems are limited, most of the time only a certain number of the most recent measurements are stored before the data are periodically overwritten. This causes only the current data to be available.

Data that concern more than just a single system are passed on through paper documents. These documents are usually batch supply notes, trial orders, etc.

Due to this structure, larger-scale analyses (e.g. the entire production cycle or historical measurements) are subject to the following limitations:

- Current and historical data are not always available.
- The data are incorrect (e.g. due to incorrect measurements, input errors or manipulation).
- Data might be missing as a result of process management, due either to data loss (periodic overwriting) or lost paper documents.
- The data are spread over various storage media, a phenomenon which often leads to serious media disruption.

Figure 1: In production facilities with older industrial machines, in most cases data are only available locally, for a limited time or in paper format.
These machine-specific problems are also accompanied by IT issues:
- The data are in completely unstructured data pools and their numerous structures are different and fixed.
- The systems are not always networked.
- Most of the time, data are also simultaneously kept in completely different applications:
  - Excel solutions for spreadsheets and macros
  - Access applications with a wide range of data sources
  - Rigid storage structures as parameters for storing documents/data

The frequent lack of consistency also leads to problems caused by different operating systems and versions of standard applications. Additionally, there is often a lack of standard interfaces and consistent user guidance.

The use of existing data must therefore be kept far behind the technical possibilities as there can be no systematic evaluation. Consequently, it is necessary to identify existing problems and, following a concept phase, to make the data available for analysis.

3 The route to information: collecting and structuring

An increasing number of machines can be networked with one another thanks to technical “innovations”. Even database systems, which previously were only meant for ERP systems, are seeing more and more frequent use in production. These allow measurement and production data to be stored securely and transparently. This is the basis of the first steps towards networked production.

The initial target is a structured database; this requires existing data to be systematically consolidated. This is already supported in a modern production environment as the systems are supplied with modern data structures.

With older systems, however, the following steps must first be taken:
- A database system consolidates the various individual pieces of data from the parallel data storage.
- Database applications replace Excel and paper documents.
- Various automatic functions connect programmable logic controllers with the database. Other systems are integrated simultaneously.
The overall structure converts the existing data into information, although not all systems have to be integrated in the first step. Even if, initially, just a data pool is created for selected systems, this still guarantees high data availability that has a holistic structure and is up to date. Defined by the IT structure, the relevance of the data depends on the intervals at which they are imported from the data source into the database.

These “information islands” are networked to form one large entity in several stages (see figure 2). This can be done, for example, by means of a company-wide data centre or by merging existing database systems, and allows for ad hoc evaluations across entire production chains. The company can define the scope of data to be merged for each system separately. The only condition is that the informational value of the individual data is not lost; these are the very base data required for future evaluations.
4 Analysis of the current situation: an application for initial success

As soon as the data are in a structured data pool, the current situation can be analysed using custom applications which have been adapted to the specific ways the user thinks and works. This way, a planner with no in-depth IT knowledge can very quickly draw conclusions about the process or product directly from their work station.

As the data are up to date, free from errors and more accessible, the production process can be analysed over an extended period of time. Additionally, identical process steps or machines can be compared with one another, even in real-time if the data are sufficiently up to date.

A specific example: The transparent mapping of the entire process chain makes it possible to immediately intervene in the production process if a problem occurs. The number of flawless parts produced can be optimised throughout the entire process if defective parts are identified and separated quickly. Separating defective parts as quickly as possible improves quality as well as efficiency because unnecessary work steps on NOK parts are ultimately minimised.

Moreover, targeted data storage broadens the range of quality management measures and product tracking: this makes it possible to react quickly if quality or even field defects occur. Specific partial blocks or recalls will then contribute to only selected parts from a series having to be recalled instead of entire batches or daily productions.

The up-to-date data that have been collected can also be used to calculate other key figures (see figure 3):

- OEE figures including availability, performance and quality factors
- Process and machine capability
- Real-time inventories
- Current inventory between various processes

![Figure 3: Mapping and comparison of various processes in terms of their merchandise value and inventory](image)
Continuous analyses of the current situation lead to larger volumes of data which can then be used for more in-depth analyses. To this end, the data have to be transformed from a production database into an analysis database. Depending on the purpose of the analysis, there are various ways of processing the data:

- Online analytical processing (OLAP) is ideal for quickly analysing large volumes of data or testing hypotheses.
- Statistical analyses produce new knowledge about the internal structures of data.
- Learning and regression algorithms make forecasts about future measurements.

### 5.1 Online Analytical Processing (OLAP)

The purpose of OLAP is to quickly evaluate hypotheses. The data material is prepared in the form of a multi-dimensional cube. The results of the queries serve to confirm or refute hypotheses. As the hypothesis determines the structure of the cube, prior knowledge of the desired purpose is required and has to be taken into consideration. A new multi-dimensional cube might be required for other analyses.

### 5.2 Data Mining

Another type of analysis deals with finding new links by examining the data material for previously unknown relations. This does not require any prior knowledge of the desired purpose.
Possible algorithms:
- **Correlation analysis**: The calculation of correlations between two different features determines dependencies. This makes it possible, for example, to quickly identify and prevent future errors.
- **Scatter plots**: a simple comparison of characteristics (see figure 4)
- **Clusters, classification**: these identify two or multi-dimensional links, fields or groups within the data pool (see figure 5).

### 5.3 Learning and regression algorithms

The final type of analysis is designed to predict the values of other characteristics on the basis of a few characteristics. At the start of production, for example, measurement data can be used to make predictions about future measurements, which in turn can be used to improve the management of the process.

Algorithms:
- **Decision trees**: Automatically generated sequential query lists can limit potential target results.
- **Neural networks (artificial intelligence)**: Predictions can be made about future events on the basis of automatically identified dependencies within a volume of data.
- **Regression analysis**: Predictions about certain characteristics can be made by means of approximation through functions (see figure 6).

This shows that OLAP and data mining really can derive new knowledge from information. Below are some examples of possible new results and what they can be used for:

- Predictive maintenance intervals
- Various characteristic clusters
- Meteorological influences on production
- Determination of the quality of machines over extended periods
■ Frequency of faulty deliveries
■ etc.

The fields of application, possible uses and results for OLAP and data mining are enormously diverse and therefore have to be adjusted individually.

6 Expansions to the data structures

Data management can be changed as soon as the advantages of the in-house production data pool have been exhausted in a factory such as with cloud solutions. Even the content can be expanded, leading in this case to a data warehouse.

6.1 Cloud solutions

The shifting of data management to a public cloud makes it possible to analyse the entire supply chain – that is, if the partners make their data available. This produces completely new production management options. Overall, not only does this option lead to “larger-scale” data management, but it can also serve as the basis for new business models. In many sectors this market has yet to be explored or shaped, which is why cloud solutions should always be taken into consideration.

Figure 6 (CC-BY-SA Sigbert): Quadratic regression model
6.2 Data warehouse

One alternative is to expand the in-house database. Company data such as orders and resources can broaden the data pool, which will no longer consist solely of production data. This total integration of all data will lead to a holistic data warehouse – namely, a complete and thorough company database – and facilitate the analysis and optimisation of all business processes.

7 Other examples and big data outside of production

The high availability of data opens up many other opportunities to use data profitably. Here are some examples:

- **ERP and resource planning:** The data can be used to improve the utilisation and efficient use of resources.
- **Sales analyses:** Identification and improvement of sales options and the optimisation of sales, logistics and planning.
- **Supply chain and ALM:** The integration of the supply chain can improve quality and efficiency depending on the process environment; if the product life cycle on-site with the customer is taken into account in the chain, additional lucrative service fields such as service, maintenance and recycling can be opened up.
- **Maintenance and support on the premises of the customer:** Tracking a product and collecting data from the customer can decrease the frequency of breakdowns as downtime can be accurately predicted by analysing data (predictive maintenance). This is particularly relevant with regard to expensive technical equipment such as aeroplanes, medical devices and lifts.
- **New business models through data analysis:** Lend instead of sell, give away and make money from additional services.
8 Summary

Especially in older manufacturing companies, improved data management leads to improved efficiency through analyses of the current situation and data mining. The specific analyses and procedures to be used can be freely defined by the customer. The basis can be established without setting any fixed targets and additional requirements can be added at a later date. The rough procedure is as follows:

- Digitise, standardise and store data
- Network data through systems
- Use analyses of the current situation and data mining to generate knowledge and optimise processes
- Identify and integrate data that have still gone unused

In coordination with its customers, infoteam Software AG was quick to dedicate itself to developing database applications. A wide range of data sources (PLC data, documents, Oracle databases, MS SQL Server) have been processed using a wide range of technologies (Java, C#) in numerous different projects. In these projects, we have analysed vertical and horizontal processes for our customers and optimised and improved them with a range of software solutions.

9 Glossary

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALM</td>
<td>Application Lifecycle Management</td>
</tr>
<tr>
<td>ERP</td>
<td>Enterprise Resource Planning</td>
</tr>
<tr>
<td>Field defects</td>
<td>Field defects are product defects which first occur when the product is being operated by the end customer; this can lead to expensive recalls.</td>
</tr>
<tr>
<td>NOK</td>
<td>Not OK</td>
</tr>
<tr>
<td>OEE</td>
<td>Overall Equipment Effectiveness</td>
</tr>
<tr>
<td>OLAP</td>
<td>Online Analytical Processing</td>
</tr>
<tr>
<td>PLC</td>
<td>Programmable Logic Controller</td>
</tr>
</tbody>
</table>
Contact

infoteam Software AG
Am Bauhof 9
D-91088 Bubenreuth
Telefon: +49 (0) 9131 / 78 00 - 0
Telefax: +49 (0) 9131 / 78 00 - 50
info@infoteam.de
www.infoteam.de

infoteam Software AG
Emil-Figge-Straße 80
D-44227 Dortmund
Telefon: +49 (0) 231 / 97 42 56 - 0
Telefax: +49 (0) 231 / 97 42 56 - 09
dortmund@infoteam.de
www.infoteam.de

infoteam Software AG
Eichtalstrasse 55
CH-8634 Hombrechtikon
Telefon: +41 (0) 44 927 15 15
Telefax: +41 (0) 44 927 15 10
info@infoteam-software.ch
www.infoteam-software.ch

infoteam Software (Beijing) Co., Ltd.
Zhongguancun North Street 151
Yan Yuan Resource Tower, Room 820
CN-100080, Haidian District Beijing
Telefon: +86 (0) 10 5887 6786
Telefax: +86 (0) 10 5887 6785
info@infoteam.com.cn
www.infoteam.com.cn