

Applying Big Data Concepts to Improve Flat Steel Production Processes



Jens Brandenburger, Valentina Colla, Silvia Cateni, Antonella Vignali, Floriano Ferro, Christoph Schirm and Josef Melcher

Abstract In this chapter we present some results of the first European research project dealing with the utilisation of Big Data ideas and concepts in the Steel Industry. In the first part, it motivates the definition of a multi-scale data representation over multiple production stages. This data model is capable to synchronize high-resolution (HR) measuring data gathered along the whole flat steel production chain. In the second part, a realization of this concept as a three-tier software architecture including a web-service for a standardized data access is described and some implementation details are given. Finally, two industrial demonstration applications are presented in detail to explain the full potential of this concept and to prove that it is operationally applicable. In the first application, we realized an instant interactive data visualisation enabling the in-coil aggregation of millions of quality and process measures within seconds. In the second application, we used the simple and fast HR data access to realize a refined cause-and-effect analysis.

J. Brandenburger (✉)

VDEh-Betriebsforschungsinstitut GmbH, BFI, Düsseldorf, Germany

e-mail: jens.brandenburger@bfi.de

V. Colla · S. Cateni · A. Vignali

Scuola Superiore Sant'Anna, SSSA, Pisa, Italy

e-mail: colla@sssup.it

S. Cateni

e-mail: s.cateni@sssup.it

A. Vignali

e-mail: a.vignali@sssup.it

F. Ferro

ILVA S.P.a, Novi Ligure, Italy

e-mail: floriano.ferro@gruppoilva.com

C. Schirm · J. Melcher

Thyssenkrupp Rasselstein GmbH, Andernach, Germany

e-mail: christoph.schirm@thyssenkrupp.com

J. Melcher

e-mail: josef.melcher@thyssenkrupp.com

© Springer Nature Singapore Pte Ltd. 2018

S. S. Roy et al. (eds.), *Big Data in Engineering Applications*,

Studies in Big Data 44, https://doi.org/10.1007/978-981-10-8476-8_1

Keywords Big data • Manufacturing • Flat steel production
Data visualization • Analytics

1 Introduction

According to the German ICT industry association BITKOM.

Big Data means the analysis of large amounts of data from a variety of sources at high speed aiming to generate economic benefit [1].

In other words, the aim of Big-Data is not to create vast data pools, but to make use of the data in a target-oriented manner.

From the viewpoint of manufacturing industries, this means that it is much more important how to combine the data with production knowledge than to process huge amounts of data in real-time. In this context, the focus must be set on the usability of Big-Data and not only on the technological limits of data processing. Therefore, in [2] the concept of the domain expert is introduced. Contrary to the technological expert, he does not care about the well-known Vs,¹ but demands 3 Fs to Big-Data applications:

For the domain expert the usage of Big Data should be Fast, Flexible and Focused.

Thus, in this chapter we present a solution that tries to maximize data usability, already before storage, by means of a multi-scale data representation. Accompanied with the knowledge about the production history of each single product this can be used to implement ETL-procedures providing tailored data for any kind of through process analysis, enabling fast, flexible and focused Big-Data applications.

2 Problem Definition

Today modern measuring systems support the production of high quality steel and provide an increasing amount of high resolution (HR) quality and process data along the whole flat steel production chain. Although this amount appears to be quite small compared to the huge amount of data that occurs e.g. in the world wide web, especially the complexity of the flat steel production chain makes detailed analytic tasks difficult for classical relational database management systems (RDBMS).

¹Volume, Variety, Velocity, Veracity

2.1 Spatial Querying

First reason impeding data analytic tasks in flat steel production is the fact, that the product (steel coil) basically implies a 2-dimensional spatial object. Whereas on the one hand it is no problem for a standard RDBMS to access the whole HR data for each single coil produced quite fast, it is much more difficult to retrieve only partial data.

If the task is:

Retrieve all available HR data within a certain coil-region over one whole production period containing thousands of coils.

Standard data-warehouse concepts are approaching their limits very fast. To formulate spatial queries using standard SQL leads to extremely complex statements. Spatial extensions of common RDBMS like MS SQL Server 2008 [3] or postGIS for PostgreSQL [4] are able to cope with spatial objects, apply spatial indices and execute spatial queries, but they are devoted to supply geographical data.

For instance in case of the MS SQL Server the spatial indices are based on regular grid structures in different resolutions approximating spatial objects [3]. The main problem of this approach is that the regular grid structure should preferably be small compared to each single spatial object described by this index to perform well. However, in case of single surface defects measured by an automatic surface inspection system (ASIS) this relation is just the opposite. There are many small objects detected on the coil that would need a very small grid structure for the spatial index and thus cannot be efficiently covered. The underlying index structure is dedicated to less and larger objects as usually the case in geographic applications but not for industrial data.

The addressing and aggregation of a multiple-coil request, as foreseen for HR data evaluation is not supported directly and thus not of high-performance. Figure 1 shows the result of a trial setup using a geospatial database (PostgreSQL + PostGIS) and GeoServer [5] as frontend, which is an open source geospatial web service engine dedicated to process geographical data. The applicability and performance of this architecture was tested by means of Apache JMeter [6] measuring the querying and aggregating performance of surface defects of 5 coils at a time (different for each request).

In this trial, an average response time of 550 ms on the basis of 1500 requests was obtained, but the response time increased exponentially with the number of queried coils, leading to unacceptable results if more coils are analysed [7].

However, for quality monitoring and improvement of the production processes a more statistical view on the data is mandatory. The absolute coil-position of a measurement loses importance and a normalized view on the data should be used to compare not only single coils but also full production cycles and/or material groups regarding suspicious data distributions. This involves not only 5 but several

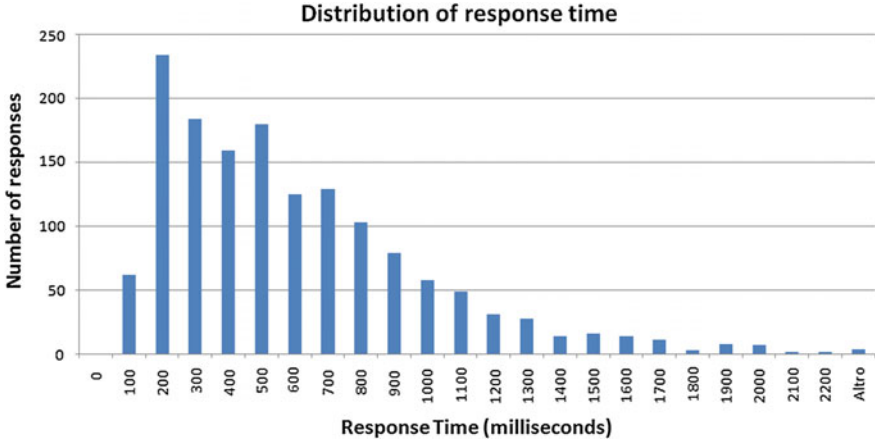


Fig. 1 Distribution of response time for spatial querying the data of 5 coils

hundred coils thus a new concept of HR data representation had to be found allowing the fast aggregation of spatial information over a massive amount of individual coils.

2.2 Product Tracking

To be able to locate a quality problem in the flat steel production chain it is mandatory to be able to track each individual coil or part of a coil through the full production process. As exemplary shown in Fig. 2 for the tinplate production, such a production chain can be quite complex.

Moreover, it is very likely that an effect of a quality problem is measured at a process step different from the step that causes that problem. Consider an ASIS installed at a tinning line located at the end of the processing chain that detects surface defects emerging at a degreasing line.

To be able to correlate process parameters with such a quality problem, it must be possible to calculate the emerging position p^* of each single surface defect from the position p of the ASIS detection at the tinning line as

$$p^* = (T_{d \rightarrow t}^c)^{-1}(p) \quad (1)$$

where $T_{d \rightarrow t}^c$ is the individual position transformation of coil c from the degreasing line to the tinning line. This transformation is a composition of the following base transformations occurring in the flat steel production process:

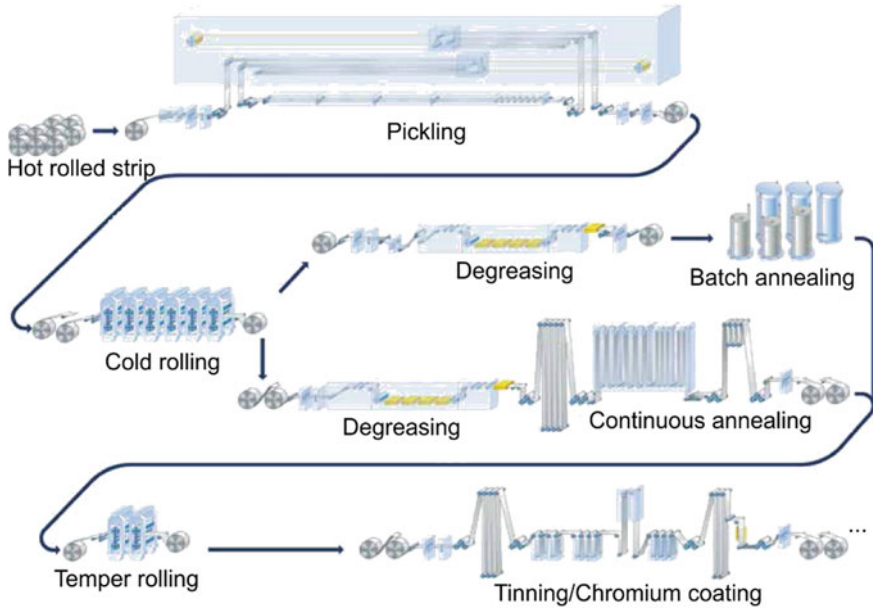


Fig. 2 Exemplary production chain for tinplate production [8]

- **winding**—always causes a switch of coil start and end. Furthermore a switch of top/bottom and left/right side is possible
- **production process**—rolling processes cause coil elongation and thus linear position shifts
- **cutting/welding**—due to continuous production, customizing and repairing operations coils may be cut and new coils assembled from multiple coil parts of preceding process steps

In Fig. 3 each blue dot on the most right picture represents the relative position of one or more surface defects as detected by an ASIS installed at the tinning line. These defect positions are transformed to the preceding lines according to the available tracking information for the individual coil. Consequently, the quality data of each coil has to be transformed individually before they can be combined with process data from preceding lines making this kind of through-process investigations very complex and time-consuming for standard data-warehouse concepts. A dedicated storage concept should be able to consider tracking information already in the ETL-procedures to provide fast query response times for data analytic tasks.

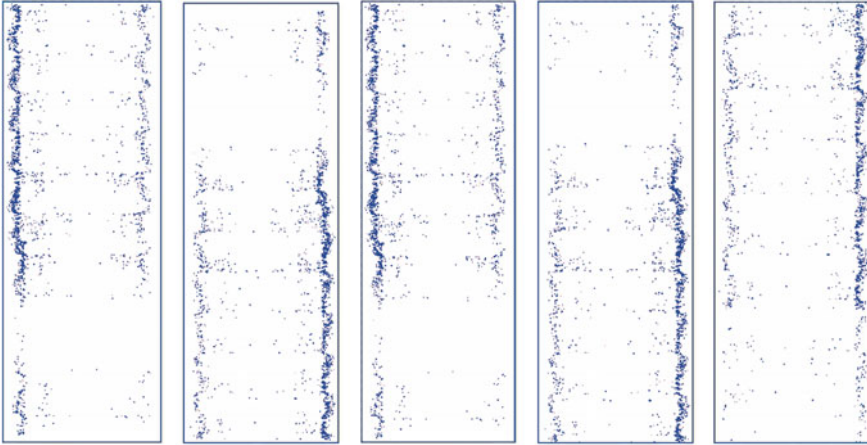


Fig. 3 Position of surface defects tracked through the production chain

3 How to Apply Big Data Concepts in Flat Steel Production?

The HR data that occurs in flat steel production are values provided by measuring systems installed at different steps in the production chain. Thus for a suitable production model, each processing step of a single product has to be considered as one product instance.

Furthermore, all gathered data has to be assignable to an individual product and each measurement to a single position on this product. This means in particular, that time-series must be synchronized with the production and consequently all data becomes spatial. To allow synchronization between different production steps additionally the full tracking information of each coil should be available. Figure 4 shows the different data types that have to be considered as HR-data in flat steel production.

Usually for further processing this kind of information is aggregated based on constant length segments (e.g. 1 m) and stored in a factory-wide quality database (QDB). Thus, already today common applications based on the available data in a QDB can monitor the current production, support quality decisions or allow dedicated investigations in case of customer claims [9]. However, to realize a solution supporting as well fast in-coil aggregations as displaying data from the viewpoint of different production steps by considering material tracking information, a tailored data model is needed.

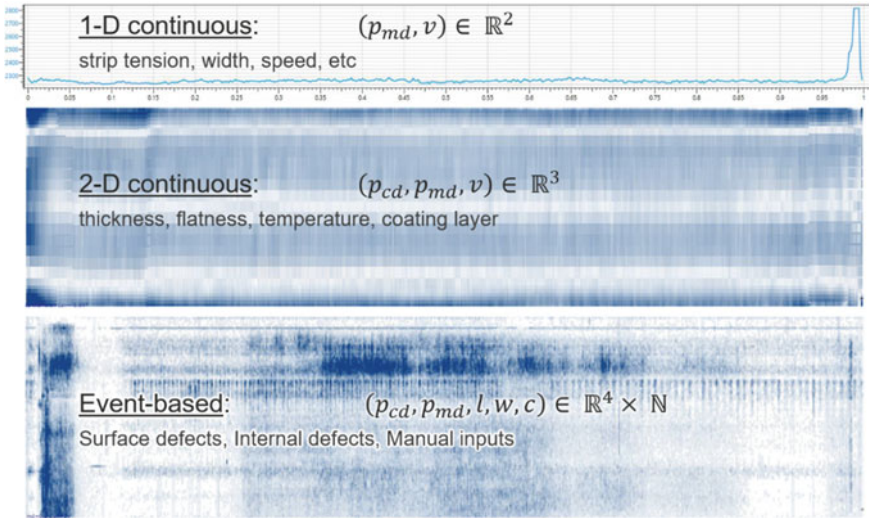


Fig. 4 Example of measurement data entries in flat steel production, with coil length-position p_{md} , coil width-position p_{cd} and measurement value v or event attributes: length l , width w and class c

3.1 Production Data Model

A suitable data model that is able to provide efficient HR data access for flat steel production was developed in the European research project ‘EvalHD’ [7]. The general idea is to address the problem from the domain expert point of view. To visualize HR-data on the screen an image has to be created representing a pixel matrix $I := [0, N_x] \times [0, N_y]$. Each pixel $p_{x,y} \in I$ represents a rectangle $R_{x,y}$ on the coil that is located relative to each pixel position in I [10]. The color (resp. value) of each pixel is calculated from the HR measurements within the corresponding rectangle $R_{x,y}$.

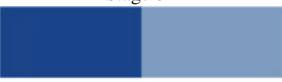
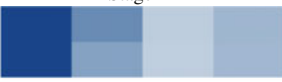




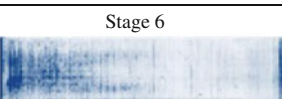
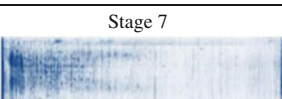
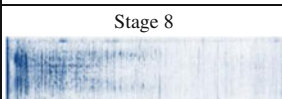
Thus, for visualisation and a given image I it is sufficient to store only the position x, y together with the aggregated value $p_{x,y}$ in a production data model without losing any information. By means of a bijective function

$$\mu: [0, N_x] \times [0, N_y] \rightarrow [0, N_x N_y] \quad (2)$$

furthermore a unique $TileID := \mu(x, y) \in [0, N_x N_y]$ can be assigned to each pixel position simplifying the aggregation over multiple coils significantly as this aggregation can be directly performed over equal $TileIDs$.

Moreover, this means that using a grid structure fitted to the size and resolution of the intended visualisation I is a minimal data representation as it stores exactly the data that is required. On the other hand there are some applications where a resolution much lower than $N_x \times N_y$ is reasonable. One example is the

Table 1 Visualisation of surface defect distribution over about 8000 coils in different resolution stages

Stage 0  n = 2	Stage 1  n = 8	Stage 2  n = 32
Stage 3  n = 128	Stage 4  n = 512	Stage 5  n = 2048
Stage 6  n = 8192	Stage 7  n = 32768	Stage 8  n = 131072

cause-and-effect analysis described in 5.2, but also for visualisation it may be beneficial to use lower resolutions. They can be used to realize a ‘coarsest first’ visualisation and a user-experience similar to other modern rendering engines as applied e.g. by the virtual globe “Google Earth” and in detail described in [11]. A parallel querying of the desired information over multiple resolutions and the immediate visualisation of the finest available data as soon as it is completely processed leads to a low response time and high user-acceptance of the system [12].

This idea leads to a multi-scale grid representation of measurement data as shown in Table 1. According to Eq. (2) in this representation each grid cell can be uniquely addressed by means of the pair $(Stage, TileID)$, thus keeping the fast aggregation capabilities of the single stage model.

For the final production data model a grid of 1 cell in cross direction (CD) times 2 cells in production direction (MD) was chosen as coarsest stage 0 resolution. The different number of cells was chosen because of the unequal length to width ratio of a steel coil (often $< 1:10000$). For the next stage, the resolution is multiplied by 2 in each dimension leading to the final grid hierarchy shown in Table 2. In this setup, exactly 4 grid cells of stage $i + 1$ fit in a grid cell of stage i and thus each $TileID$ of stage $i + 1$ can be uniquely assigned to a single stage i grid cell.

To extract the raw data from the productive databases, transform and load them into the common HR data model at first each coil has to be normalized to a length and width of 1. This means that each point $P_c := (x, y)$ on a coil c is converted to a new point

Table 2 Grid definitions and exemplary sizes of grid cells for a coil length of 7500 m and a coil width of 1500 mm

Stage	Tiles CD	Tiles MD	$\Delta x_{c,s}$ ($c_w := 1500$ mm) (mm)	$\Delta y_{c,s}$ ($c_l := 7500$ m) (m)
0	1	2	1500	3750
1	2	4	750	1875
2	4	8	375	937.5
3	8	16	187.5	468.75
4	16	32	93.75	234.38
5	32	64	46.88	117.19
6	64	128	23.44	58.59
7	128	256	11.72	29.3
8	256	512	5.86	14.65

$$\overline{P}_c := \left(\frac{x}{c_w}, \frac{y}{c_l} \right) \in [0, 1]^2, \text{ with } c_w := \text{coil width and } c_l := \text{coil length.} \quad (3)$$

Consequently, the reachable synchronization accuracy can be calculated dependent on the coil c , the coil dimensions c_w , c_l and the resolution stage $s \in [0, 8]$ as

$$\Delta x_{c,s} = \frac{c_w}{s} \text{ resp } \Delta y_{c,s} = \frac{c_l}{s} \quad (4)$$

Some exemplary values for $\Delta x_{c,s}$ and $\Delta y_{c,s}$ are also given in Table 2.

Once each coil position is normalized, the transformation of point-based raw data into the grid structure can be performed quite easy by simple cell-based aggregation of all measurements falling into one specific grid cell. Regarding 1D and 2D continuous measurements, the aggregations stored in the grid structure are minimum, maximum, mean and count of the measuring values. Event-based data (like surface defects) are usually stored as rectangular regions combined with a certain identifier describing the type of the event (e.g. defect class) and can either be aggregated as absolute counts per grid cell or overlapping area relative to the full cell area.

Given this multi-grid data representation, the question remains how to enable the combination of data across production stages. This again can be easily solved by not only simultaneous storage of data across different resolutions, but also across different perspectives. Assumed that the information about all coil transformations is given during data transformation, the data can be tracked upstream and/or downstream and further grid data can be created and stored for each measurement from the perspective of other production steps. The data is stored for each plant separately according to the available material tracking information. Thus, the data is available simultaneously in different plant coordinates enabling fast HR data access by means of redundant data storage.

Finally, to analyse production data and find causes of quality problems it is essential to be able to filter data according to different production parameters, like material, thickness, production period, etc. Thus, further filter conditions have to be added to each grid entry of the same HR-type (see Fig. 4) to keep filter capabilities of the data representation. The grid attributes finally stored can be classified into five different categories:

- **Coil Filter**—Unique coil identifier that allows filtering grid entries by coil attributes like material type, thickness, process parameters, etc.
- **Identifier**—Unique grid cell identifier needed for fast aggregation (Stage, TileID)
- **Sub Filter**—Further type specific filter conditions (defect class, measuring device, etc.)
- **Data**—Per grid cell aggregated measuring data (min, max, mean, count)

This production data model is able to synchronize and aggregate HR data of different kind from different perspectives very fast. Therefore, it acts as a kind of database index on the available HR raw data supporting dedicated querying of grid data.

4 Implementation

The production data model described above was implemented as classical three-tier architecture as shown in Fig. 5. This architecture has the benefit that it separates presentation, application processing, and data management functionality.

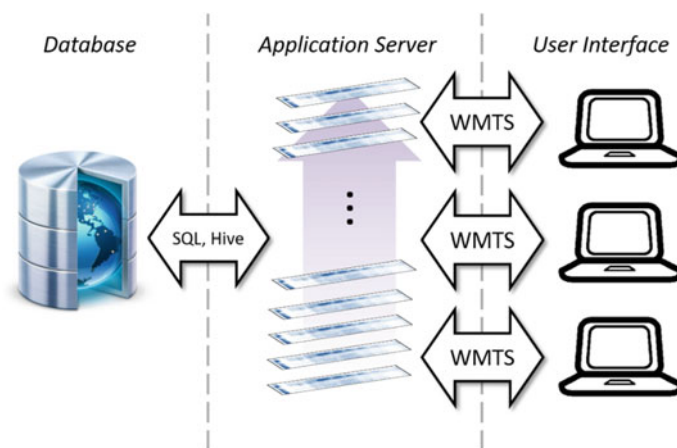


Fig. 5 Schema for HR data access

At the bottom of this architecture, a database management system (DBMS) implements the HR-data model. In this approach, it is not relevant if the database is a standard RDBMS or a Hadoop cluster. The application server has to cope with it and use the correct query syntax to provide the desired grid data by means of parallel querying the employed database. On top of this architecture, a browser application communicates with the application server following a unified web-service definition that is based on the Web-Map-Tile-Service (WMTS) standard provided by the Open-Geospatial Consortium [13]. In the implemented setup the querying of the data follows a two-step approach:

1. Query all coils meeting certain filter conditions applied by the user
2. Query grid data according to the selected coils

The resulting grid data can be provided either aggregated (for visualisation) or per-piece (for cause-and-effect analysis). If material tracking should be considered one important detail of the final implementation is, that each coil queried in the first step knows its own production history. This allows switching the viewpoint to another process step without re-querying the selected coil-set. Furthermore, it is possible to select only coils that were processed at a certain line being another important aspect when searching for quality problem causes. For further details on the web-service definition, please refer to [14].

5 Application

To proof the usability of the architecture depicted in Fig. 5 it was finally implemented at two industrial sites. The production data was transformed to grid data and continuously imported in the HR data model. Based on the available data a solution for the fast data visualisation was realized supporting instant-interactive data analysis and a solution for refined cause-and-effect analysis was implemented.

5.1 Visualisation

The system implemented at thyssenkrupp Rasselstein GmbH finally involved 1137 HR-measurements from 24 main aggregates of the complete tinplate production chain together with the full material tracking information. This includes data from the hot strip mill to the finishing lines at the end of the production. As database, an MS SQL Server 2012 has been chosen with a capacity of 20 TB being sufficient to cover approx. 1 year of full production grid data.

It was necessary to put a lot of effort in the implementation of the import services to be able to store the available HR data to the server without flooding. Extensive use of methods like bulk inserts, parallel processing and index-free temporary tables

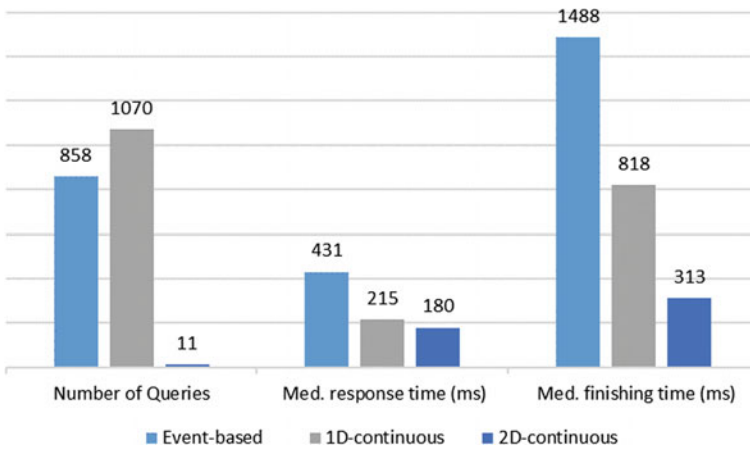


Fig. 6 Performance statistics of HR server over 2 months usage

were required to finally achieve ‘coil-realtime’, meaning that the time required for data storage can follow the production. It can be reasonable assumed that this will be no issue using a database system dedicated to Big-Data processing. On the other hand, it has to be investigated if the query performance of such a system can be competitive with the index structures provided by the standard RDBMS.

Figure 6 shows a performance statistic over two months of system usage. In this period the median response time of the system, providing defect data was 215 ms. This response time refers to the first visualisation of the lowest resolution stage queried. The querying process was implemented by means of parallel SQL-queries for 8 equally sized full width stripes distributed over the full coil length. In this trial the multi-scale visualisation started with stage 2 and refined over stage 6 before finally stage 8 results were presented.

On average (median) the browser application was able to provide the full resolution defect data to the user (8 stripes at stage 8 resolution) in less than 1.5 s. Furthermore, it can be seen that the usage of the system in the testing period was mainly focused on the analysis of 1D and event-based data, whereas 2D-continuous data played only a minor role.

5.1.1 Paw-Scratch Example

The following example clearly demonstrates the benefit of the developed solution as it shows how a quality problem could be successfully solved using the interactive visualisation solution presented in this chapter. The quality problem investigated was the so-called ‘paw-scratch’—defect that often looks like a paw print of an animal. This defect is well detected by ASIS and can be classified very reliably by using context information in post-processing rules [9]. Thus, it is a good choice for

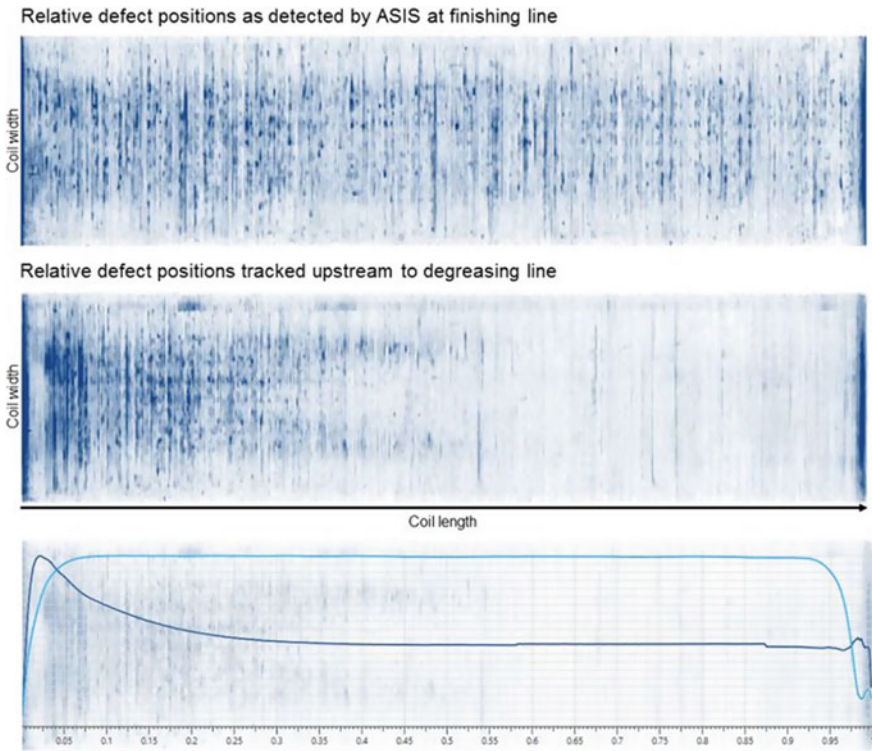


Fig. 7 Relative defect positions of paw-scratch defects at the finishing (top) and tracked to the degreasing line (middle). Bottom: in-coil aggregated mean values of related process variables (light: line speed, dark: strip tension)

a detailed ASIS data analysis as no manual data verification is required [15]. The investigation started with the analysis of paw-scratch defects as detected by an ASIS installed at the finishing line. The visualisation on top of Fig. 7 shows the distribution of this type of defect over a set of more than 2000 coils affected by this defect and combines more than 500.000 single defects in one image. Herein the most blue grid positions represent more than 500 single paw-scratch detections.

The picture in the middle of Fig. 7 shows the same result as the top picture after each single ASIS result of each individual coil has been tracked to one of the two degreasing lines located at the thyssenkrupp site in Andernach. In this case, a characteristic distribution of the paw-scratch defects becomes visible and it appears that significant more paw-scratches were located at the beginning of the coils.

This example impressively shows what happens if no tracking information is considered for data analysis. Due to the individual coil transformations as described in Sect. 2.2, the characteristic defect distribution at the causative line gets completely lost throughout the production chain. Thus, in this case no reasonable

correlation analysis can be performed by means of the defect positions as measured by the ASIS.

Generally, an almost uniform distribution as shown on top of Fig. 7 is a strong indication that the cause of this defect is not located at the specific line. On the other hand once there is a kind of characteristic distribution visible as in the middle, it makes sense to correlate the relative positions of the surface defects with process parameters of the plant.

The bottom picture of Fig. 7 shows an overlay of the paw-scratch distribution at the degreasing line and the stage 8 grid mean values of the 1-dimensional process data 'line speed' (light) and 'strip tension' (dark). It can be seen that there is a strong correlation between the dark strip tension graph and the paw-scratch locations.

For the quality engineers this correlation was taken as reason to perform some trials, how to adapt the process parameters of the degreasing line in such a way that a lower strip tension at the beginning of the coil can be achieved. To evaluate the trial results again the ASIS data of the finishing lines had to be tracked back to the degreasing line to see if the new control strategy led to the desired result. This again could be done pretty easy using the developed visualisation system.

Thus, iteratively the paw-scratch problem could be solved and today 90% less coils are affected by paw-scratches than before the implementation of this system.

5.2 *Cause-and-Effect Analysis*

The occurrence of ripple defects in the course of the Hot Dip Galvanizing (HDG) process on thick coils (i.e. thickness ≥ 1.5 mm) with low zinc coating (i.e. in the range 50–71 g/m²) has been examined at ILVA s.p.a. Ripples are vertical line shaped defects that could be designed as diffuse coating ruffles so that they are identified by ASIS systems without difficulty at the end of HDG lines [16]. The process parameters, which mainly affect the occurrence and the significance of ripples, are the air blades configuration, cooling techniques, process speed and wiping medium. The real effects of each process variable deviation is still not very clear; skilled personnel control ripple defects by employing nitrogen as wiping medium in air blades but it is not always an effective method and this action could increase costs uselessly. Moreover, a greater understanding of the phenomenon under observation can improve the quality by decreasing reworked or scrapped material.

The above-described problem has been dealt by analysing data from a HDG line at ILVA, including 1D-continuous HR measurements of 20 process attributes that can be categorised into four categories:

1. Air blades;
2. Temperatures (zones before and after the zinc bath, top-roll, water bath);
3. Line speed;
4. Fan coolers.

The case study has been divided into two analysis considering the use of nitrogen in the air blades. The first analysis is devoted to study the process conditions that minimize the ripple presence despite only air is blown and the second analysis regards the knowledge of process conditions that lead to a high defectiveness while employing Nitrogen. This analysis is important as the Nitrogen is expensive and it is interesting to minimize its use maximizing its effectiveness.

When only air is blown, the target is to find a set of process conditions that allows minimizing the ripple occurrence, while, on the other side, when nitrogen is employed, the target is to avoid the occurrence of ripples at all. Nitrogen is in facts expensive, thus, its use should be minimized and its effectiveness maximized.

Due to this reason, two datasets were organized for air and nitrogen blowing, respectively, which comprised HR measurements of the process attributes highlighted above as inputs, and a binary classification of the tiles (null value for tiles without ripple defect and unitary values for highly defective tiles) has been carried out. Dataset is composed by about 360 coils that are developed through the HR data model dealt in 3.1 and pre-processed in order to remove outliers performing a multivariate Fuzzy-based method (FUCOD) that is in detail designated in [17, 18].

Another issue regards the fact that the available variables are 1D-continuous, while defects are 2D-continuous. In order to aggregate, input and target tiles are combined into so-called ‘slices’ along the coil width by summing ripple defects along that direction. An example of the stage 2 slices is shown in Fig. 8.

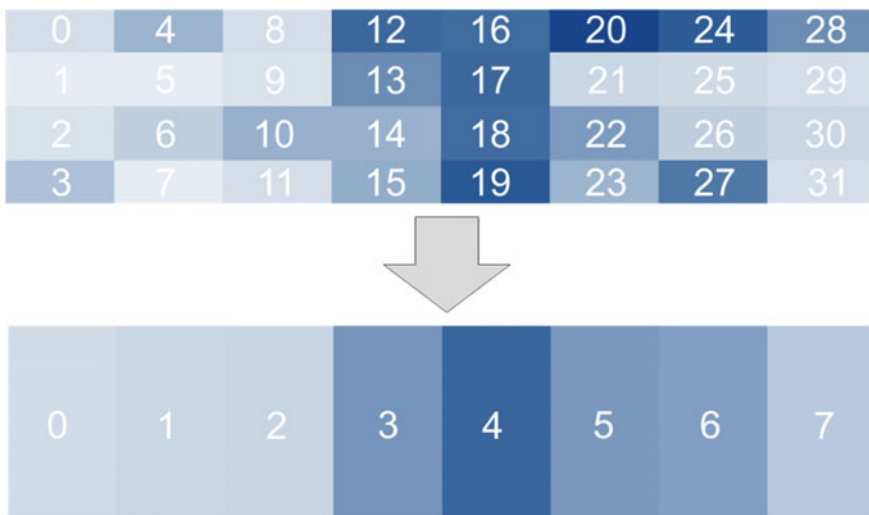


Fig. 8 Tile aggregation along coil width (top: tiles for stage 2, bottom: aggregation of tiles to create associated slices)

A binary classification based on Decision Tree (DT) has been developed; class 0 represents the slices with the total absence of defects, while class 1 identifies highly defective slices. With the term *highly defective*, we indicate slices with a number of defects that exceeds a threshold. The threshold is automatically computed and fixed to the 95th quantile of the empirical cumulative distribution of the percentage area of defects.

Dataset is randomly shuffled and a training and a validation set are defined preserving the initial proportion among the two classes. The training include the 75% of the available samples while the validation set is composed by the remaining 25%. For both case studies (air and Nitrogen blowing), classifier based on Decision Tree (DT) has been carried out and subsequently validated on the respective validation set [19].

The performance has been evaluated computing the Balanced Classification Rate (BCR) as defined in Eq. 5.

$$\text{BCR} = \frac{1}{2} \left[\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right] \quad (5)$$

where TP is the number of unitary values correctly classified, FN the number of unitary values incorrectly classified, TN the number of null values correctly classified, and FP the number of null values incorrectly classified. BCR is more appropriate for imbalanced datasets than the classical accuracy index, as in both available datasets the null class is far more frequent [20–22].

Each node of the trained DT represents the associated process variable, each branch corresponds to a range of values it can assume and finally the leafs correspond to the two defined classes. Through a path from the root to a leaf the procedure detects a process window leading to a specific result, taking into account if the leaf value is unitary or null.

Decision tree classifier can be translated in a simple chain of IF-THEN-ELSE rules becoming easily interpretable by no-skilled operators. This method provides an actual way to support decisions and to extract good process windows to be adopted during production to avoid defects; moreover, it can be adopted to provide a degree (namely *importance*) of how much a process variable affects the analysed target, so that the quality experts can further investigate on it [23].

Table 3 illustrates the very satisfactory performances of the classifier, while Table 4 shows the selected variables that mostly affect the target for the two

Table 3 Classification performances in terms of BCR evaluated on the validation set

Case study	BCR	Accuracy class 0	Accuracy class 1
Air-blowing (%)	99.34	99.78	98.90
N2-blowing (%)	97.65	98.92	96.37

Table 4 Most affecting process variables and the associated normalized relevance

Air blowing		Nitrogen blowing	
Air blade distance	1	Water bath temperature	1
Tunnel zone temperature	0.69	Air blade distance	0.46
Line speed	0.29	Hot bridle zone temperature	0.39
Air blade pression	0.25	Line speed	0.14
Fans speed	0.15	Air blade pression	0.14
Top-roll zone temperature	0.08	Fans speed	0.11
Water bath temperature	0.07	Air blade height	0.05

sub-problems, which is representative of the 95% of the information content. The proposed method is generic and do not require any a priori assumptions, for this reason it can be employed in other applications [24].

6 Comparison to Common Concepts

The described HR data model natively provides a solution for the problems of flat steel production data synchronization and material tracking, which need to be solved before a through-process data analysis can be performed in this environment.

In the present implementation, the model has been realized by means of a common RDBMS as the available amount of data allows to realize the full model data during ETL procedures and to store it completely on the server. However, this model could also be realized using Big Data Management Technologies and MapReduce to improve its scalability.

Figure 9 shows the median query response and finishing times, already mentioned in Sect. 5.1, depending on the number of queried coils. According to the linear approximation (dotted lines), it can be stated that the response time seems to be almost independent on the number of queried coils, whereas for the finishing time there is a slight increase. This behaviour is mainly achieved due to the shift of the tracking consideration and the data synchronization to the ETL procedures and the use of an adequate index structure on the RDBMS.

In contrast, the situation for a common data-warehouse concept, which is optimized for a fast per-coil data access, is shown in Fig. 10. It shows the finishing times for a visualisation of coil-sets of 10, 100 and 1000 coils and indicates the distribution of the finishing times measured in 5 runs for each coil-set size. Before the visualisation result can be presented, each coil has to be tracked individually and the data has to be synchronized. This leads to a clear dependency of the finishing times on the number of queried coils and a query time of approx. 10 min for a visualisation of 1000 coils.

In other words, when 1000 coils are queried, the presented data model is approx. 300 times faster than the common data-warehouse concept. This relation is even

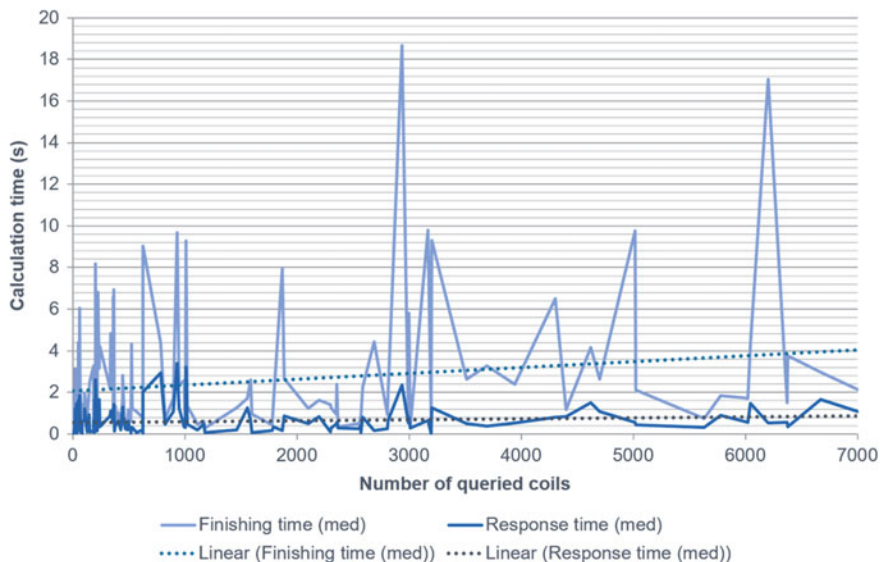


Fig. 9 Performance of the implemented visualisation solution (event-based data)

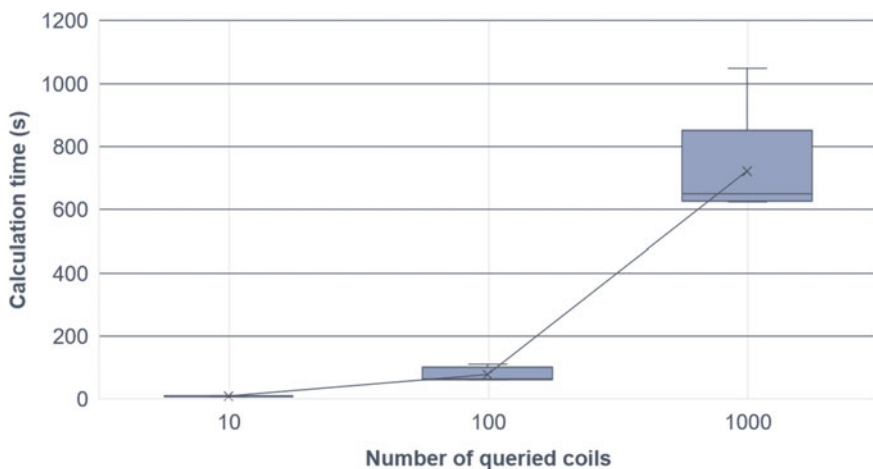


Fig. 10 Query performance of a common data-warehouse concept (event-based data)

increasing when more coils are queried making our concept outstanding fast. Thus, for the first time ever the HR query response times allow an instant interactive analysis of HR data as soon as a quality problem occurs, enabling a new dimension of quality data assessment for flat steel production [14].

7 Conclusion

For an effective application of Big-Data technologies in manufacturing industries, it is not sufficient to store a massive amount of raw data. Instead, a full production model is mandatory to enable through process synchronization of all available measuring data.

Therefore, in this chapter we describe a suitable production model for flat steel production, able to realize fast, flexible and focused access to industrial Big-Data, due to a new multi-scale data representation across production steps. Using a three-tier architecture, we could successfully implement this approach at two industrial sites and proof its usability. Moreover, as well for a fast data visualisation supporting the interactive investigation of quality problems as for providing source data for HR cause-and-effect analysis using more than one aggregated value per coil, we could show that this new concept performs far better than any state-of-the-art data model in terms of query response time.

Concluding it can be stated that standard data-warehouse concepts are not appropriate to utilise the full potential of modern measuring equipment in flat steel production, as an efficient statistical evaluation of multi-coil HR data is not adequately supported. On the other hand, new technologies combined with a suitable production model can provide valuable input to quality engineers and plant operators already from the very beginning.

Acknowledgements The work described in the present paper was developed within the project entitled “Refinement of flat steel quality assessment by evaluation of high-resolution process and product data—EvalHD” (Contract No RFSR-CR-2012-00040) that has received funding from the Research Fund for Coal and Steel of the European Union. The sole responsibility of the issues treated in the present chapter lies with the authors; the Commission is not responsible for any use that may be made of the information contained therein.

References

1. Bartel, J., Decker, B., Falkenberg, G., Guzek, R., Janata, S., Keil, T. et al. (2012). Big Data im Praxiseinsatz: Szenarien, Beispiele, Effekte. Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. (BITKOM).
2. Freytag, J.-C. (2014). Grundlagen und Visionen großer Forschungsfragen im Bereich Big Data. *Informatik-Spektrum*, 37, 97–104.
3. Katibah, E., & Stojic, M. (2011). New Spatial Features in SQL Server Code-Named ‘Denali’. SQL Server Technical Article. <https://msdn.microsoft.com/en-us/library/hh377580.aspx>.
4. PostGIS. <http://postgis.net/>.
5. GeoServer. <http://www.geoserver.org>.
6. Apache JMeter. <https://jmeter.apache.org/>.
7. Brandenburger, J., Schirm, C., Melcher, J., Ferro, F., Colla, V., Ucci, A., et al. (2016). Refinement of Flat Steel Quality Assessment by Evaluation of High-Resolution Process and Product Data (EvalHD). European Commission, Directorate-General for Research and Innovation.
8. thyssenkrupp Rasselstein. (2015). Wege der Produktion. Brochure.

9. Brandenburger, J., Piancaldini, R., Talamini, D., Ferro, F., Schirm, C., Nörtersheuser, M., et al. (2014). *Improved Monitoring and Control of Flat Steel Surface Quality and Production Performance by Utilisation of Results from Automatic Surface Inspection Systems (SISCON)*. European Commission, Directorate-General for Research and Innovation.
10. Brandenburger, J., Schirm, C., & Melcher, J. (2016). Instant interactive analysis—how visualisation can help to improve product quality. In *Surface Inspection Summit SIS*. Europe, Aachen.
11. Tanner, C. C., Migdal, C. J., & Jones, M. T. (1998). The Clipmap: A virtual Mipmap. In: *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 151–158). ACM.
12. Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann Publishers Inc.
13. OGC OpenGIS. (2010). *Web Map Tile Service Implementation Standard*. Open Geospatial Consortium Inc.
14. Brandenburger, J., Colla, V., Nastasi, G., Ferro, F., Schirm, C., & Melcher, J. (2016). Big data solution for quality monitoring and improvement on flat steel production. In *7th IFAC Symposium on Control, Optimization and Automation in Mining, Mineral and Metal Processing MMM*, Vienna.
15. Brandenburger, J., Stolzenberg, M., Ferro, F., Alvarez, J. D.; Pratomongo, G., & Piancaldini, R. (2012) *Improved Utilisation of the Results from Automatic Surface Inspection Systems (IRSIS)*. European Commission, Directorate-General for Research and Innovation.
16. Borselli, A., Colla, V., Vannucci, M., & Veroli, M. (2010). A fuzzy inference system applied to defect detection in flat steel production. In *IEEE World Congress on Computational Intelligence, WCCI*.
17. Cateni, S., Colla, V., & Nastasi, G. (2013). A multivariate fuzzy system applied for outliers detection. *Journal of Intelligent and Fuzzy Systems*, 24, 889–903.
18. Cateni, S., Colla, V., & Vannucci, M. (2007) A fuzzy logic-based method for outliers detection. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications* (pp. 561–566).
19. Cateni, S., Colla, V., & Vannucci, M. (2010). Variable selection through genetic algorithms for classification purposes. In: *Proceedings of the 10th IASTED International Conference on Artificial Intelligence and Applications, AIA* (pp. 6–11).
20. Vannucci, M., Colla, V., Cateni, S., & Sgarbi, M. (2011). Artificial intelligence techniques for unbalanced datasets in real world classification tasks. In: *Computational Modeling and Simulation of Intellect: Current State and Future Perspectives* (pp. 551–565).
21. Vannucci, M., & Colla, V. (2011). Novel classification method for sensitive problems and uneven datasets based on neural networks and fuzzy logic. *Applied Soft Computing Journal*, 11, 2383–2390.
22. Vannucci, M., & Colla, V. (2015). Artificial intelligence based techniques for rare patterns detection in the industrial field Smart Innovation. *Systems and Technologies*, 39, 627–636.
23. Cateni, S., Colla, V., & Vannucci, M. (2009). General purpose input variables extraction: A genetic algorithm based procedure GIVE a GAP. In *ISDA 2009—9th International Conference on Intelligent Systems Design and Applications* (pp. 1278–1283).
24. Cateni, S., Colla, V., Vignali, A., & Brandenburger, J. (2017). Cause and effect analysis in a real industrial context: study of a particular application devoted to quality improvement. In *WIRN 2017, 27th Italian Workshop on Neural Networks June 14–16*, Vietri sul Mare, Salerno, Italy.