# Data Provisioning for Service and Warranty Analytics


Authors

**Mario Palmer-Huke**
Director Analytics, EXA GmbH

**Oliver Merten**
Head of Business Intelligence, EXA GmbH

# Content

# Introduction

According to many studies the hot topic in Business Intelligence for 2016 is "Predictive Analytics".[1] While over the last years many companies have invested in Proof-of-Concepts (POCs) and trials, it now seems to be the time for them to invest into productive solutions. According to a BARC study the percentage of German companies planning investments in Predictive Analytics solutions has risen to 44%[2].

However, in the same BARC study the same companies named "Integrating Heterogeneous Data Sources" as their most pressing need (score of 3,66 on a scale of 1 (lowest) to 5 (highest)).

In our opinion both results are closely related to each other and reflect our own experience with numerous clients. Almost since the inception of EXA, we have been working with clients on implementing predictive solutions. And almost every time the biggest hurdle in getting these programs off the ground was providing a correct and comprehensive data basis for the desired analytical use cases. In that sense, introducing advanced analytics today is not so much different from introducing better BI capabilities over the last decade.

The following document outlines our experience in providing the necessary data for analytics in a SAP HANA architecture. It focuses on the data acquisition, cleaning and storage.

# Some Boundary Conditions

EXA AG provides IT Services with special focus on Analytics on SAP Platform for the Manufacturing Industry vertical. EXA has done a few projects for its manufacturing clients that focus on advanced analytics in the field of Service and Warranty Analytics, which is described by a recent study as the clear number #1 use case in this industry[3].

The focus of this whitepaper is on how to provide the necessary data integration framework for Service and Warranty Analytics on SAP HANA or a similar Platform. The reason our clients have chosen SAP HANA are diversified and include:

- Performance requirements
- HANA being part of their overall architecture
- Leverage  in-house experience in managing and developing SAP solutions

Though most of our learnings in the following pages would apply to any advanced analytics solution for Service and Warranty Analytics, many of the technical best practices and recommendations focus on HANA specifically and would need to be adjusted for other technology platforms.

[1] For instance: IBM Institute for Business Value: Analytics – The upside of disruption, 2015
[2] Source: BARC Research Study: Modernes Datenmanagement für die Analytik, 2015
[3] Source: Manufacturing and the data conundrum: Too much? Too little? Or just right? - An Economist Intelligence Unit Report, 2015

# Typical Data Subjects

For projects in the field of Service and Warranty analytics, companies typically use 4 different and distinct kinds of data:

- Machine generated / sensor data
- Master Data
- Transactional data from OLTP systems (e.g. ERP, CRM, ...)
- Other internal and external data (unstructured, weather, D&B, ...)

## Machine Generated & Sensor Data

This type of data usually generates very large volumes. It is not uncommon for a sensor to generate up to billions of records per year. This data is generally transmitted in one of the following manners:

1. *Summary transmission:* The device or machine sends a summary set of relevant data points to a central system. Often latency is in minutes or hours and the relevance of data is already determined at the source. A typical example is telematics data in the vehicle industry or a medical device which transmits usage-related or defect data. The use cases typically do not involve real-time monitoring and alerts but are focused on gaining additional insights beyond operational efficiency.
2. *Continuous transmission:* A set of sensors continuously sends its data back to a central system. The relevancy of the data (individually or collectively) is only determined at the time of storage or after it has been stored. Data arrives as a stream and usually the latency is very low. Typical examples include real-time monitoring of devices for failure or quality predictions.

As per our experience, any machine data used for business scenarios in the field of Service and Warranty Analytics fall into one of these 4 categories:

- *Simple sensor readings*: e.g. "oil temperature", "driver out of seat" - status of a specific machine/device characteristic
- *Calculated*: e.g. "# of accelerations that lasted longer than 3 seconds and ended at a vehicle speed above 140km/h", "device operating within specifications for X number of days"- usually representing some kind of complex machine status. Data is commonly not transmitted real-time.
- *Alerts*: specific message sent once a certain threshold is exceeded
- *Environmental conditions*: e.g. "temperature", "soil pH-value" - sensor readings of the environment in which machine/device is operating. These can be single readings or complex combinations.

## Master Data

For Service and Warranty Analytics, the majority of master data required is around machine/product and customer/partner dimensions (e.g. type of machine installed at building x of client location y).

The data is transferred from the system of records or a specialized master data system into the data acquisition layer. In that sense the data can come from ERP and/or CRM or any other system where such data is maintained.

We have found that many times, new master data sets need to be defined and created. If the business process, for instance, is changed from selling a device to selling a service, installation locations of such devices become relevant for correct billings as the final customer might request to bill by site and not send an invoice per single device. Existing, available master data may not have such groupings yet.

It is very rare that master data from external agencies is acquired. Examples would include but are not limited to geospatial data or very specialized data attributes for clients, dealers and partners.

## Transactional Data

Significant value is achieved by bringing together the previously disconnected low-level, machine related data with operational business processes, like service cases and warranty claims.

The additional transactional data is supposed to be analysed side-by-side with the machine provided data. Here, the data subjects are completely driven by the use case and differ widely. Often valuable data is not accessible in a structured form as for example detailed information of a warranty claim might only be available as scanned text pages or free text entered in a warranty system. In addition, availability and access to systems with the requested data might not to be setup.

## Other Internal and External Data

Data in this category is use case specific. Basically it describes data that is necessary to get an end-to-end view. If for instance the temperature and humidity in which a device operates influence failures, joining the GPS date/time with external weather records might be the way to go. If for a new service based business model, billing entities need to be verified, sourcing Dun & Bradstreet data could be a solution.

However, the data will either be used as additional master data or provides another form of transactions. It can be both - structured or unstructured. Therefore, we do not describe specific challenges for this type of data as possible solutions can be found within the described examples.

# Common Challenges in Acquisition, Fusion and Storage of the Data

As outlined earlier, we consider multiple data subject areas relevant for Service and Warranty Analytics. Based on our project experience, we have elaborated typical challenges and possible solutions below -

# The Challenges with Machine Generated (Sensor) Data

Typically, we face 2 areas of concern with machine data in each project – data quality and data volume & performance challenges.

## *Data Quality*

Normally, one would assume that data quality issues are not overly present with machine generated data. On the contrary, we believe that there are some common and actually large issues that one has to deal with in every project.

Common challenges include:

- *Dropped and missing data*: it is not uncommon that data records just do not arrive due to errors in the transmission (e.g. issues with mobile phone connection or records not even considered for transmission for various reasons). In such cases, one has to decide if to inter- or extrapolate the missing data. Initially this is a business decision and both answers ('yes' and 'no') can make sense. In case data needs to be "created", it needs to be decided where and how that should happen. The necessary functionality is available in HANA. Based on the data size and frequency, streaming solutions like Sybase ESP (Event-Streaming-Processing) are used in conjunction with HANA. The Smart Data Streaming feature of HANA can also be used to process small amounts of data in very short time, thus they can be used to decide which data is usable "as-is", which data will be dropped and which data needs to be processed further.
- *Wrong data:* sensors can provide wrong data or data can get falsified during transmission. In any case corrections need to be done when loading the data. These corrections are normally applied during the ETL process. To give an example: in one particular case sensor data needed to be equidistant (meaning for instance 1 reading of the current temperature every second). Unfortunately, it did arrive with missing data points and hence those missing data points had to be interpolated during the data load process.
- *Joining data:* in an ideal world keys are unique and data can be joined straight away. In reality however, data structures change over time so it can happen that what based on business logic should deliver a unique single record (answer), provides multiple returns (e.g. each physical machine should be only at one location at a time or only 1 item per header should exist). There are various reasons for such scenarios:
    - Data structures/business rules change over time: Especially when historical data is processed, the data does not necessarily fit to business rules that are valid today.
    - Data from multiple sources doesn't have a common governance: Although different sources might use the same key fields (e.g. master data and machine data might both have a machine id as unique identifier) there's no central governance, i.e. it could happen that – although logically not feasible, but technically possible – the same machine ID identifies different machines.

It should be noted, that addressing data quality challenges might lead to volume and performance challenges. Hence it is important to invest time in planning each specific case where such processing shall take place. In our example of the missing sensor readings, the inter-/extrapolation could have

been done in HANA or in a pre-processor. In one such case, since we had high load and report performance requirements in HANA, we decided to use an external tiering concept – like described below – and did the data processing outside of HANA. This way we could spread the processing load and finalize the equidistant temperature data set before finally loading it into HANA.

## *Volume & Performance*

One of the biggest challenges with machine generated data is the volume and performance aspect. The challenges can be addressed internally via dynamic tiering or partitioning, or by creating an external hot/warm/cold-architecture.

### Internal Approach

As sensors can create billions of records in a few months, tables in HANA become very large. SAP recommends partitioning at around 2 billion records, but our experience shows that table query and processing performance starts to degrade notably at 1b+ records.

In our opinion there are 2 distinct options to deal with the situation -

1. *Dynamic Tiering:* For multi-temperature memory strategy, the extended table concept with SAP HANA Dynamic Tiering can be used. Using this concept, you can store warm data - with specific usage profiles without functional restrictions - in extended tables, which are managed by SAP HANA. This enables to optimize usage of the main memory in SAP HANA. It depends on multiple criteria (for instance total data volume) if an approach within HANA is the optimal and preferred solution. In projects where a *Dynamic Tearing* approach is not feasible, we recommend using Sybase IQ / Hadoop in the architecture next to HANA as depicted further below.

2. *Partitioning:* a good approach if the final user queries/use cases can be anticipated with a high certainty. If that is the case, a good partitioning strategy can be defined upfront. However, based on project experience, we strongly recommend to do so based on thorough analysis of the expected work-loads. Furthermore, the strategy should be validated with thorough performance tests to verify the results. We also recommend, to re-run a performance regression test scenario after each new HANA release to verify the performance improvement results are still valid.

### External Approach

If Dynamic Tiering does not address the volume challenges sufficiently, a similar concept can be built off-loading warm/cold data into another data structure like Sybase SQ or Hadoop.

In contrast to the HANA Dynamic Tiering approach, this might lead to some functionality restrictions for immediate data usage. On the other hand, this approach also allows for (pre)-processing of data outside HANA to, for instance, interpolate missing data points which then can be provided to HANA. Once the dataset is finalized it could be loaded into HANA or accessed through SDA.

When building an overall big data architecture, the off-HANA data storage could act as a data repository and serve multiple other analytic solutions (“data lake” concept).

### *Other Considerations*

For all analytical purposes it is recommended to align all data to the same time zone. That can be UTC or any other local time zone. But sometimes the question is how to determine the correct source time? For fix deployed machines and sensors the location is known which also means the corresponding time zone can be easily determined. However, for moving devices (e.g. planes, vehicles, ...) etc. it is not necessarily a fixed attribute. Depending on the kind of transmission different strategies can be applied:

- *Summary Transmission:* The time zone logic can be applied either by the device itself (when preparing the data) based on some logic like GPS, device time etc. or the time-stamp can be recalculated on the receiver side using a similar logic.
- *Continuous Transmission*: In these situations, the sensor usually does not have the computing capabilities for attaching the right time nor can it be transmitted. Take the example of a temperature sensor. The sensor might transmit temperature, if it is advanced, maybe also an additional device ID. In that case the time can only be attached once the data is received. Ideally this can be made part of the CEP Engine, like SAP ESP (Event Stream Processing) or HANA Smart Data Streaming.

## The Challenges with Master Data

When looking at Master Data following 3 points need attention:

1. Access and quality of the required Master Data
2. Need for new Master Data
3. Not being able to apply/use Master Data to the business use case problem

For the first point, the challenges with Master Data for Service and Warranty Analytics are not different from the challenges with Master Data in any other BI or DWH project and hence can be addressed in the same manner: change the source, create based on other data or implement a workflow requesting manual intervention. Incorrectly maintained customer addresses are the classical example for this case.

The solution for the second point is as straight forward: the newly required Master Data needs to be created or stored in a place where it can be accessed and used within HANA. New Master Data is usually linked to a particular use case. For example, if analytics are based on location, geospatial data is needed. Or it might be relevant to model client company structure beyond just name/address in order to be able to send invoices to different business departments.

The last point is the most difficult one. It might for instance be necessary to group machines based on certain attributes that are not part of the general product hierarchy (e.g. current engine firmware release). In this case the following approaches should be considered:

- Semi-automatic creation of hierarchies or links between data elements that will enable the intended analytic use cases. To stay with the firmware release example, based on machine data received, the actual release version is identified and vehicles are grouped accordingly.
- Build or find relationships by applying text search functionality. By browsing the text, relationships between items and groups/clusters can be identified. By using fuzzy search such connec-

tions can be proposed with a certain probability and should be verified and confirmed by users. For later (re)use confirmed or falsified relations can be stored in the database. The applied technology is similar to what we describe below in the section about "using text" and "other unstructured data".

We used this approach to group vehicles based on relations between alerts they had generated and specific free-text warranty claims. The created vehicle groups could then be used for further cluster analysis.

## Using Text and other Unstructured Data (other Transaction Data)

One of the most common challenges for Service and Warranty Analytics is to process text and link it to structured data. In a typical use case involving warranty claims, those will usually be filed as unstructured text while all the alerts or sensor data are machine readable. So how can these data elements be linked for further insights?

One possible solution is to load the texts as text fields (block) into HANA and initially index them using HANA Text Analysis capabilities. After this initial step, the following logic can be applied:

- One can limit the possible records by looking at date dependencies. Simply put, any alert after a warranty claim was filed is probably not the root cause while any alert that was raised shortly before a machine breakdown might be a pointer to it. Therefore, the business logic needs to be customized to each project.
- Alerts contain text descriptions that can be part of the transmitted alert. Alternatively, the alert is created in the ESP or HANA based on a combination of sensor status. Either way we can use the HANA based text search feature to find links. The quality of the link can be improved by using fuzzy search and configure the fuzzy parameters based on some trial and error (the parameters have to be tweaked and adjusted to provide the best results for a given use case). The result set can then be used for analytics directly or given to a person via workflow for further improvements.

## General ETL Considerations

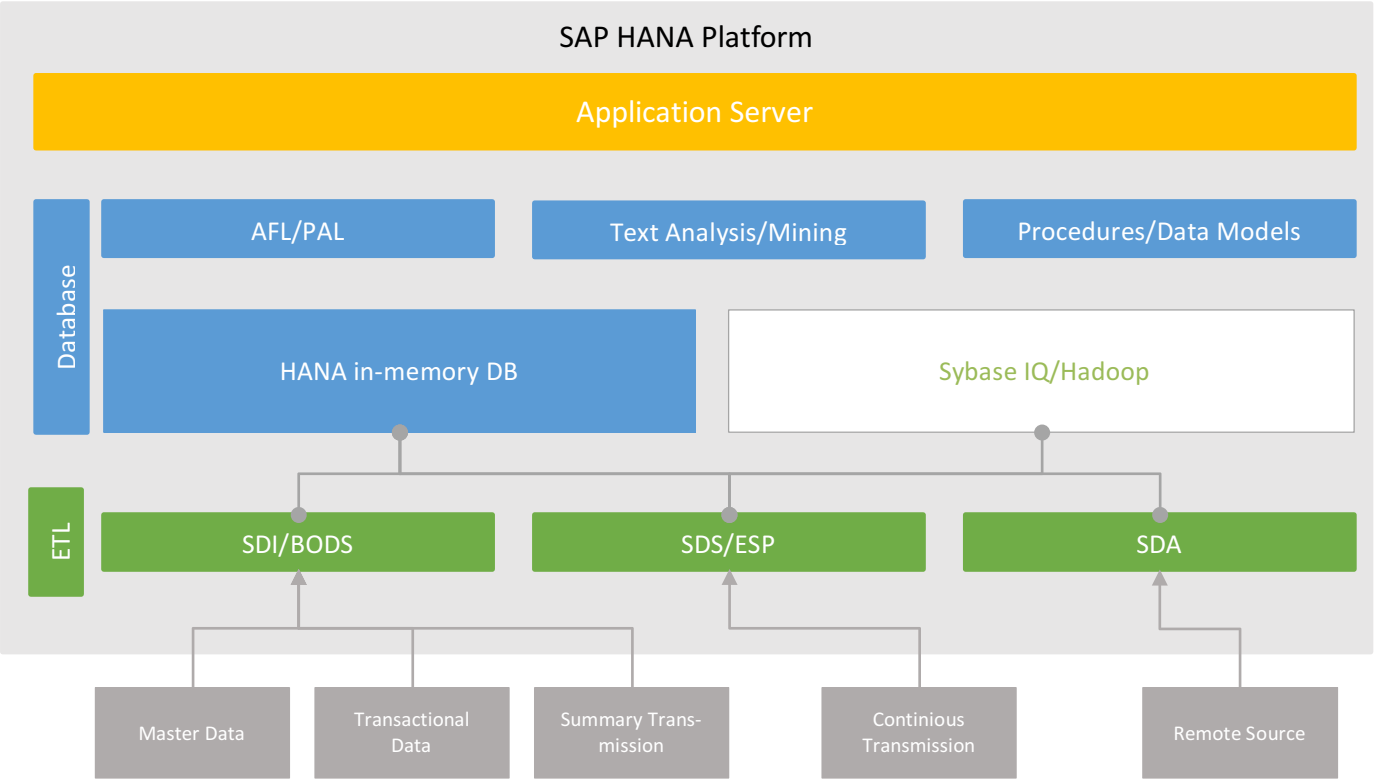
For any data load in HANA – as a matter of fact for any load into any analytical database - we recommend using an ETL solution and not to rely on scripts or similar solutions. Via an ETL tool the project can ensure better reliability, availability, security und auditability. Also, a potential porting of the load procedures to another target DB or host system will be much easier.

Accepting this logic, in our opinion, leaves 2 scenarios for Service and Warranty Analytics in the SAP world: using a stand-alone ETL tool like BO Data Services or the native HANA Smart Data Integration (SDI) framework.

ETL tools provide more flexibility for targets (beyond HANA) and a larger functionality set. On the other hand, HANA SDI reduces the infrastructure foot-print and potentially overall licensing cost and lastly provides better HANA integration.

# Reference Architecture

Taking all of the above into considerations, EXA has developed a HANA-based reference architecture for using structured, unstructured, machine and system data in analytical uses cases.




*Picture 1 - Schematic Reference Architecture*

The modularization of the architecture allows for deploying only the parts necessary for the initial use cases. At the same time, it is prepared and able to grow with the need for more diverse business scenarios.

# Use Cases for Service and Warranty Analytics

As stated, we clearly see the data integration and provisioning as a key hurdle in having success with Service and Warranty Analytics. At the same time just putting data together does not give a business benefit. So beyond the hype around "predictive maintenance", what are real business challenges that companies have addressed with Service and Warranty Analytics:

- **Reducing Costs Associated with Warranty Claims:** The key is to identify issues early and also to address/fix them early. Everybody will agree that in a perfect world, a system could predict asset breakdowns based on alerts and machine data. However, most of the times it is very difficult to achieve. What one of our clients has done as an intermediate step, is to provide users

with a guided analytics system which will link alerts & sensor data from his vehicles with warranty claims filed by customers. This way patterns can be found and addressed early and in the process warranty costs are reduced significantly.

- **Building a Service based Business Model:** a company started selling service instead of machine as a business model where the customer is charged for the output of that machine and so changed the business model. A prerequisite for such a new business model is the capability to monitor the machine remotely and receive data about the output which then will be invoiced to the client. In addition, such data is used to ensure minimum down-times which directly affect the revenue stream.
- **Minimizing Repeat Repairs:** Analysis of the condition of assets together with warranty claims and former field service repair information, technician work can be optimized. System will propose additional exchange or maintenance events for a planned technician visit in order to minimize repeat repair visits and to eventually save service costs.

# Conclusions

The document above describes the typical data subjects used in Service and Warranty Analytics. It also explains typical challenges to integrate these data subjects into a common repository which can drive the advanced analytics use cases. Furthermore, we have illustrated possible solutions to these problems.

To summarize the major conclusions from our HANA projects in the field of Service and Warranty Analytics:

1. While everybody is focused on the analytics and possible use cases, data integration is the key to success. Also, most of the project efforts will be spent in preparing sufficient, quality data.
2. Every use case has its own data requirements but by and large they will come from 4 areas: Machine/Sensor Data, Master Data, Transactional Data from enterprise systems and other Internal/External Data.
3. Technical solutions for typical challenges exist within HANA or outside of HANA. Decision on which way to go will depend on the concrete situation. A flexible, holistic architecture (as depicted above) will help in addressing today's and tomorrow's requirements.
4. SAP's development for HANA is fast and rich and with every release, SAP is still adding new functionality to already existing set of functionality that allows to provide solid advanced analytics solutions. If the client is anyway using SAP platform for its systems of records and has good SAP knowledge in-house, we do recommend using HANA platform for Service and Warranty Analytics.
5. Once the data integration and data fusion challenge has been mastered, the business use cases can be "implemented". Depending on the concrete situation, this can mean implementing/using visual analytics tools like Lumira or Tableau, creating advanced analytics using statistical models or simply providing a customizable dashboard together with a workflow. Only by using the prepared data, the business benefit will be achieved.

# About EXA

EXA is a leading technology provider engaged in delivering customized and niche solutions on SAP platform for our clients. We understand industry trends and the corresponding SAP roadmap very well through our close cooperation and co-innovation with SAP in the areas of HANA, Analytics, PLM and Fiori/UI5. This enables us to act as a trustful advisor and technology partner for our clients.

EXA AG experts possess profound knowledge on the latest SAP technologies as well as know-how in the manufacturing industry vertical with a special focus on discrete manufacturing, pharmaceutical and automotive.

Operating globally, EXA offers cost efficient and tailored services for implementation, custom development and support projects. Headquartered in Germany, EXA AG also has presence across Europe, India and the USA.