Sandro Wiedmer
Schore 432
3176 Neuenegg
sandro.wiedmer@students.unibe.ch

Data Science Project

Insights from Field Data

Conceptual Design Report

05 October 2025

Abstract

Field data matters - important knowledge is hidden in operational data. Field performance therewith represents a real-world learning opportunity. Lab data is limited and not the 'real world', it helps validating assumptions (which are used for product requirements), improving designs, and preventing repeating issues. Field data closes the critical feedback loop, besides the functions working daily with field data (technical customer support, inhouse repair, ...) also for research & development.

For this project, field data from an industrial X-ray generator platform is used. Those devices are running at end user sites located around the world, in different industries as e.g. automotive, aerospace, electronics, electric vehicles batteries and semiconductor manufacturing. This X-ray generator platform consists of a collection of product variants: There are unipolar (160 and 225kV output voltage) and bipolar X-ray generators (320 and 450kV), and for each voltage class several X-ray tube types which can be connected, optimized e.g. for cycling or continuous mode, different focal spot sizes etc.

The generators as the data source write diagnostic reports, and a part of these reports is uploaded to a cloud platform. This data is available in Databricks, where the analysis is done and models are built.

There are insights which in this project are tried to be collected from the present field data, by data analysis, descriptive statistics, questions that can be answered by statistical tests, predicting lifetime-related change of characteristics and uncover patterns and anomalies.

Table of Contents

1 Project Objectives

Having the field data coming from a number of generator devices available in Databricks, represents a significant opportunity. There are a multitude of topics which could be analyzed and help validating assumptions and improving designs. For this project just a few of them are selected.

One specific topic of relevance is arcs - electrical discharges - in the vacuum insulated X-ray tube, which is a stressful situation for the generator and influences its lifetime. Therefore, as a goal, the question to be answered by a statistical test, from the existing field data, is whether the number of arcs per exposure time and generator device is below a certain threshold.

Another topic to be analyzed from the field data using supervised learning is the degradation of a component resp. subsystem by tracking of input-output characteristics drifting over time. To be more specific, one actual issue is drifting of the X-ray tube heater current needed per emission current and high voltage operating point, observed with a part of the X-ray tubes. The goal here would be, to check the accuracy of explaining the degradation from the existing data (change in characteristics as a function of age, exposure time and number of cycles), to afterwards do a prediction of lifetime related degradation. This is a regression task, as a sub-category of supervised learning.

Unsupervised learning could be used to try to uncover patterns and anomalies in the field data. First, it is interesting to do clustering of end user operation patterns. Second, regarding anomalies, having from time to time a handful of devices where one (manually and/or semi-automated) could have a closer look, would be helpful for detecting problematic devices before a defect has arisen.

2 Methods

For the analysis in this project, the Databricks infrastructure will be used, since the field data is already ingested there and it offers a wide range of well-suited functionality. Databricks is a unified cloud platform for efficient data analysis and machine learning [1]. It handles large data volumes (e.g. as here log files resp. diagnostic reports). In Databricks, the analysisand model building can be done using Python notebooks, dashboards and SQL queries, or a combination of them.

Databricks is built on top of Apache Spark [2]. Within Databricks, Spark enables distributed data processing across scalable clusters of compute resources. Data is typically handled using PySpark DataFrames, which support efficient, parallelized operations on large datasets that exceed single-machine memory limits. For smaller or exploratory workloads, these PySpark DataFrames [3] can be converted to Pandas DataFrames, allowing more flexible local analysis

with familiar Python libraries. This hybrid workflow combines Spark's scalability with Pandas' ease of use for targeted data exploration and visualization.

Statistical test

Regarding the statistical test to answer whether the number of tube arcs per exposure time and generator device remains below a certain threshold, following test types will be used:

- Fleet-level: Exact Poisson test
- Device-level: Negative binomial regression with exposure offset (generalization of Poisson regression, for overdispersed distributions, see "10 Preliminary Studies")

Regression

For the modelling of degradation of a component resp. subsystem, different regression methods can be analyzed and the performance compared, starting with the simplest one and increasing the complexity as long as the results get better.

- multivariate linear regression
- random forest (non-linear)
- neural networks (non-linear)

If the specific case with heater current drifting per high voltage and emission current operating point is taken, the heater current would be the variable to be predicted (Y) and the high voltage an emission current, ev. more to add ("age", accumulated exposure time and/or number of cycles), are the features (X).

Clustering and anomaly detection

To uncover patterns and anomalies from field data, following methods will be used and compared:

- K-means
- Gaussian mixtures
- HDBSCAN

Expected Python imports

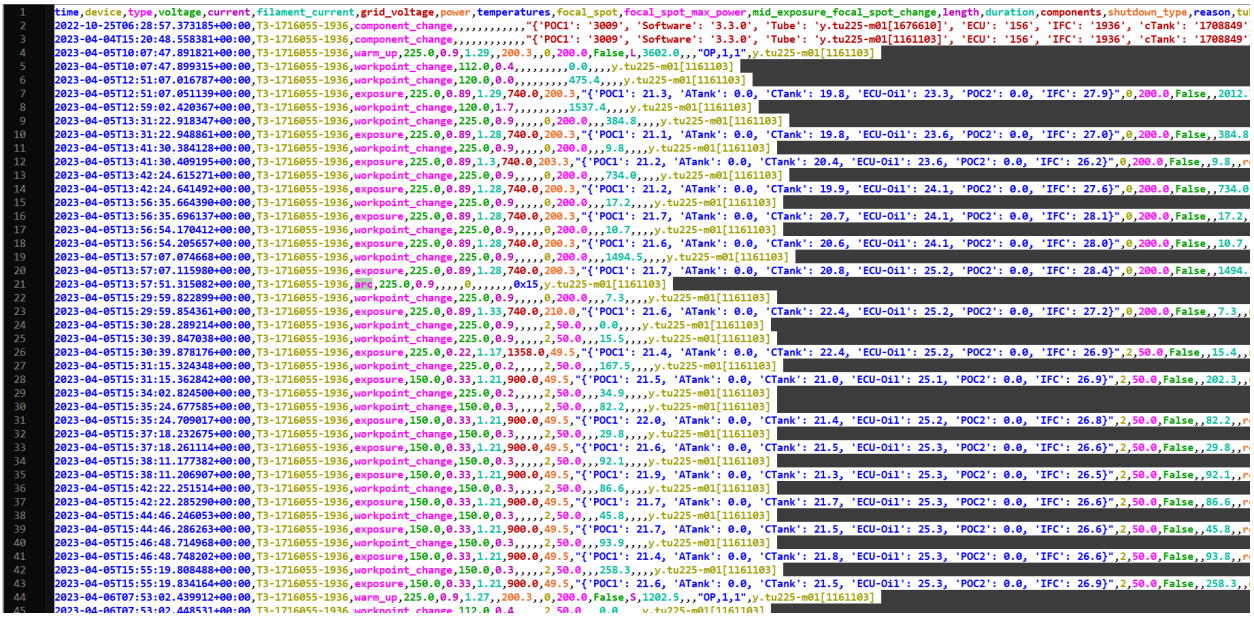
```
from pyspark.sql import DataFrame
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as st (st.poisson.cdf)
import statsmodels.api as sm (nb_model = sm.GLM(y, X,
family=sm.families.NegativeBinomial(), offset=offset)
```

```
from sklearn import linear_model
from sklearn import tree
from sklearn import ensemble
import tensorflow as tf
```

3 Data

The field data to be introduced in “7 Data Model” will be used. Currently, the number of generator devices in the field data is about 800, and will be growing while the project is being completed.

Below, an excerpt of an example diagnostic report, in the form as it can be downloaded from the AWS storage in *.csv format is shown.



```
time,device,type,voltage,current,filament_current,grid_voltage,power,temperatures,focal_spot,focal_spot_max_power,mid_exposure_focal_spot_change,length,duration,components,shutdown_type,reason,tu
2022-10-25T06:28:57.373185+00:00,T3-1716055-1936,component_change,,,,,,,,,,,,,"{'POC1': '3009', 'Software': '3.3.0', 'Tube': 'y.tu225-m01[1676610]', 'ECU': '156', 'IFC': '1936', 'cTank': '1708849'
2023-04-04T15:20:48.558381+00:00,T3-1716055-1936,component_change,,,,,,,,,,,,,"{'POC1': '3009', 'Software': '3.3.0', 'Tube': 'y.tu225-m01[1161103]', 'ECU': '156', 'IFC': '1936', 'cTank': '1708849'
2023-04-05T10:07:47.891821+00:00,T3-1716055-1936,warm_up,225.0,0.9,1.29,,200.3,,0,200.0,False,L,3602.0,,,"OP,1,1",y.tu225-m01[1161103]
2023-04-05T10:07:47.899315+00:00,T3-1716055-1936,workpoint_change,112.0,0.4,,,,,,,,0.0,,,,y.tu225-m01[1161103]
2023-04-05T12:51:07.016787+00:00,T3-1716055-1936,workpoint_change,120.0,0.0,,,,,,,,475.4,,,,y.tu225-m01[1161103]
2023-04-05T12:51:07.051139+00:00,T3-1716055-1936,exposure,225.0,0.89,1.29,740.0,200.3,"{'POC1': 21.3, 'ATank': 0.0, 'CTank': 19.8, 'ECU-Oil': 23.3, 'POC2': 0.0, 'IFC': 27.9}",0,200.0,False,,2012.
2023-04-05T12:59:02.420367+00:00,T3-1716055-1936,workpoint_change,120.0,1.7,,,,,,,,1537.4,,,,y.tu225-m01[1161103]
2023-04-05T13:31:22.918347+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,0,200.0,,,384.8,,,,y.tu225-m01[1161103]
2023-04-05T13:31:22.948861+00:00,T3-1716055-1936,exposure,225.0,0.89,1.28,740.0,200.3,"{'POC1': 21.1, 'ATank': 0.0, 'CTank': 19.8, 'ECU-Oil': 23.6, 'POC2': 0.0, 'IFC': 27.0}",0,200.0,False,,384.8
2023-04-05T13:41:30.384128+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,0,200.0,,,9.8,,,,y.tu225-m01[1161103]
2023-04-05T13:41:30.409195+00:00,T3-1716055-1936,exposure,225.0,0.89,1.3,740.0,203.3,"{'POC1': 21.2, 'ATank': 0.0, 'CTank': 20.4, 'ECU-Oil': 23.6, 'POC2': 0.0, 'IFC': 26.2}",0,200.0,False,,9.8,,r
2023-04-05T13:42:24.615271+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,0,200.0,,,734.0,,,,y.tu225-m01[1161103]
2023-04-05T13:42:24.641492+00:00,T3-1716055-1936,exposure,225.0,0.89,1.28,740.0,200.3,"{'POC1': 21.2, 'ATank': 0.0, 'CTank': 19.9, 'ECU-Oil': 24.1, 'POC2': 0.0, 'IFC': 27.6}",0,200.0,False,,734.0
2023-04-05T13:56:35.664390+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,0,200.0,,,17.2,,,,y.tu225-m01[1161103]
2023-04-05T13:56:35.696137+00:00,T3-1716055-1936,exposure,225.0,0.89,1.28,740.0,200.3,"{'POC1': 21.7, 'ATank': 0.0, 'CTank': 20.7, 'ECU-Oil': 24.1, 'POC2': 0.0, 'IFC': 28.1}",0,200.0,False,,17.2,
2023-04-05T13:56:54.170412+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,0,200.0,,,10.7,,,,y.tu225-m01[1161103]
2023-04-05T13:56:54.205657+00:00,T3-1716055-1936,exposure,225.0,0.89,1.28,740.0,200.3,"{'POC1': 21.6, 'ATank': 0.0, 'CTank': 20.6, 'ECU-Oil': 24.1, 'POC2': 0.0, 'IFC': 28.0}",0,200.0,False,,10.7,
2023-04-05T13:57:07.074668+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,0,200.0,,,1494.5,,,,y.tu225-m01[1161103]
2023-04-05T13:57:07.115980+00:00,T3-1716055-1936,exposure,225.0,0.89,1.28,740.0,200.3,"{'POC1': 21.7, 'ATank': 0.0, 'CTank': 20.8, 'ECU-Oil': 25.2, 'POC2': 0.0, 'IFC': 28.4}",0,200.0,False,,1494.
2023-04-05T13:57:51.315082+00:00,T3-1716055-1936,end,225.0,0.9,,,,,0,,,,,0x15,y.tu225-m01[1161103]
2023-04-05T15:29:59.822899+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,0,200.0,,,7.3,,,,y.tu225-m01[1161103]
2023-04-05T15:29:59.854361+00:00,T3-1716055-1936,exposure,225.0,0.89,1.33,740.0,210.0,"{'POC1': 21.6, 'ATank': 0.0, 'CTank': 22.4, 'ECU-Oil': 25.2, 'POC2': 0.0, 'IFC': 27.2}",0,200.0,False,,7.3,,
2023-04-05T15:30:28.289214+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,2,50.0,,,0.0,,,,y.tu225-m01[1161103]
2023-04-05T15:30:39.847038+00:00,T3-1716055-1936,workpoint_change,225.0,0.9,,,,,2,50.0,,,15.5,,,,y.tu225-m01[1161103]
2023-04-05T15:30:39.878176+00:00,T3-1716055-1936,exposure,225.0,0.22,1.17,1358.0,49.5,"{'POC1': 21.4, 'ATank': 0.0, 'CTank': 22.4, 'ECU-Oil': 25.2, 'POC2': 0.0, 'IFC': 26.9}",2,50.0,False,,15.4,,
2023-04-05T15:31:15.324348+00:00,T3-1716055-1936,workpoint_change,225.0,0.2,,,,,2,50.0,,,167.5,,,,y.tu225-m01[1161103]
2023-04-05T15:31:15.362842+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 21.5, 'ATank': 0.0, 'CTank': 21.0, 'ECU-Oil': 25.1, 'POC2': 0.0, 'IFC': 26.9}",2,50.0,False,,202.3,,
2023-04-05T15:34:02.824500+00:00,T3-1716055-1936,workpoint_change,225.0,0.2,,,,,2,50.0,,,34.9,,,,y.tu225-m01[1161103]
2023-04-05T15:35:24.677585+00:00,T3-1716055-1936,workpoint_change,150.0,0.3,,,,,2,50.0,,,82.2,,,,y.tu225-m01[1161103]
2023-04-05T15:35:24.709017+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 22.0, 'ATank': 0.0, 'CTank': 21.4, 'ECU-Oil': 25.2, 'POC2': 0.0, 'IFC': 26.8}",2,50.0,False,,82.2,,r
2023-04-05T15:37:18.232675+00:00,T3-1716055-1936,workpoint_change,150.0,0.3,,,,,2,50.0,,,29.8,,,,y.tu225-m01[1161103]
2023-04-05T15:37:18.261114+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 21.6, 'ATank': 0.0, 'CTank': 21.5, 'ECU-Oil': 25.3, 'POC2': 0.0, 'IFC': 26.6}",2,50.0,False,,29.8,,r
2023-04-05T15:38:11.177382+00:00,T3-1716055-1936,workpoint_change,150.0,0.3,,,,,2,50.0,,,92.1,,,,y.tu225-m01[1161103]
2023-04-05T15:38:11.206907+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 21.9, 'ATank': 0.0, 'CTank': 21.3, 'ECU-Oil': 25.3, 'POC2': 0.0, 'IFC': 26.5}",2,50.0,False,,92.1,,r
2023-04-05T15:42:22.251514+00:00,T3-1716055-1936,workpoint_change,150.0,0.3,,,,,2,50.0,,,86.6,,,,y.tu225-m01[1161103]
2023-04-05T15:42:22.285290+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 21.7, 'ATank': 0.0, 'CTank': 21.7, 'ECU-Oil': 25.3, 'POC2': 0.0, 'IFC': 26.6}",2,50.0,False,,86.6,,r
2023-04-05T15:44:46.246053+00:00,T3-1716055-1936,workpoint_change,150.0,0.3,,,,,2,50.0,,,45.8,,,,y.tu225-m01[1161103]
2023-04-05T15:44:46.286263+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 21.7, 'ATank': 0.0, 'CTank': 21.5, 'ECU-Oil': 25.3, 'POC2': 0.0, 'IFC': 26.6}",2,50.0,False,,45.8,,r
2023-04-05T15:46:48.714968+00:00,T3-1716055-1936,workpoint_change,150.0,0.3,,,,,2,50.0,,,93.9,,,,y.tu225-m01[1161103]
2023-04-05T15:46:48.748202+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 21.4, 'ATank': 0.0, 'CTank': 21.8, 'ECU-Oil': 25.3, 'POC2': 0.0, 'IFC': 26.6}",2,50.0,False,,93.8,,r
2023-04-05T15:55:19.808488+00:00,T3-1716055-1936,workpoint_change,150.0,0.3,,,,,2,50.0,,,258.3,,,,y.tu225-m01[1161103]
2023-04-05T15:55:19.834164+00:00,T3-1716055-1936,exposure,150.0,0.33,1.21,900.0,49.5,"{'POC1': 21.6, 'ATank': 0.0, 'CTank': 21.5, 'ECU-Oil': 25.3, 'POC2': 0.0, 'IFC': 26.9}",2,50.0,False,,258.3,,r
2023-04-06T07:53:02.439912+00:00,T3-1716055-1936,warm_up,225.0,0.9,1.27,,200.3,,0,200.0,False,S,1202.5,,,"OP,1,1",y.tu225-m01[1161103]
2023-04-06T07:53:02.448531+00:00,T3-1716055-1936,workpoint_change,112.0,0.4,,,,,2,50.0,,,0.0,,,,y.tu225-m01[1161103]
```

Figure 1: Example diagnostic report excerpt

The figure 2 from below shows a snippet from the corresponding table in Databricks where all the diagnostic reports are joined into one table (not all columns visible).

Sample

	time	device	type	voltage	current	filament_current	grid_voltage	power	temperatures	focal_spot
1	2024-11-14T12:15:10.800625+00:...	T3-0-1079	workpoint_change	null	null	null	null	null	null	null
2	2024-11-14T12:15:10.804841+00:...	T3-0-1079	component_chan...	null	null	null	null	null	null	null
3	2024-11-14T12:15:34.227809+00:...	T3-0-1079	workpoint_change	10.0	0.0	null	null	null	null	null
4	2024-11-14T12:15:46.192941+00:...	T3-0-1079	workpoint_change	225.0	0.0	null	null	null	null	0
5	2024-11-14T12:45:25.061408+00:...	T3-0-1079	workpoint_change	20.0	10.0	null	null	null	null	0
6	2024-11-14T12:45:28.018786+00:...	T3-0-1079	workpoint_change	22.0	10.1	null	null	null	null	0
7	2024-11-14T12:45:32.007227+00:...	T3-0-1079	workpoint_change	26.0	10.2	null	null	null	null	0
8	2024-11-14T12:45:34.011157+00:...	T3-0-1079	workpoint_change	28.0	10.2	null	null	null	null	0
9	2024-11-14T12:45:34.024268+00:...	T3-0-1079	workpoint_change	28.0	10.2	null	null	null	null	0
10	2024-11-14T12:45:35.013327+00:...	T3-0-1079	workpoint_change	29.0	10.2	null	null	null	null	0
11	2024-11-14T12:45:35.023618+00:...	T3-0-1079	workpoint_change	29.0	10.2	null	null	null	null	0
12	2024-11-14T12:45:36.064719+00:...	T3-0-1079	workpoint_change	31.0	10.3	null	null	null	null	0
13	2024-11-14T12:45:41.021864+00:...	T3-0-1079	workpoint_change	35.0	10.4	null	null	null	null	0

Figure 2: Snipped from table in Databricks, only a part of the columns is visible

The complete description of column names is shown below, under “7 Data Model”, logical data model.

For data security reasons, the customer information is separated and not accessible from Databricks. The information about which generator device ID was delivered to which customer is stored at another place. For the project here the customer per generator device is not relevant, the data from the diagnostic reports should be sufficient and therewith no security issue is expected.

4 Metadata

If reproducibility of analysis is required, as metadata, date and time of script execution is important to document, since the data changes and additional entries are written into the table(s), or freezing the tables which are being used for analysis is also possible.

Metadata for field data itself, are the units of physical values in the table, e.g. high voltage in kV, emission current in mA, durations in seconds, etc.

Another kind of metadata is customer (and location of generator device), as mentioned above under security, which is stored in other platforms, though not relevant for reproducing the analysis.

5 Data Quality

The data precision to be expected in the available field data should be more than enough, since the values come from the device internal measurements which are also used fortracking the setpoints given by the user, where high accuracy and precision is required (electronic component

tolerances, analog-to-digital conversion number of bits, input signal range etc. is chosen accordingly).

Regarding completeness it is important to mention, that the data uploaded and accessible for Databricks analysis is biased: Firstly, not all end users want or are allowed to upload diagnostic reports, e.g. because of confidentiality (no such data leaves the fab), or secondly, those end users which allow to upload data, do it mostly for devices having a problem or even defect, not for the devices which are running smoothly (herewith, accessible data is expected to have a "pessimistic" tendency).

There are many fake arcs in the data, this means an arc event is written into the diagnostic file, where in reality no tube arc happened. Also in this regard, a pessimistic view is expected, due to the number of arcs reported which is higher than what really takes place. To compensate up to a certain degree, data cleaning can be done: filter for to low voltage level tube arcs, becausethey are more realistic when reported close to the maximum specified voltage, and additionally for arc reason (from experience a part of the possible reasons are mostly fake arcs, e.g. if just a voltage drop is detected, without having an excessive current).

6 Data Flow

Data processing is organized according to a multi-layer architecture called Lakehouse Architecture with the layers Bronze-Silver-Gold (medaillon), which is best practice in Databricks [4]. The following figure shows a simplified overview of data flow from source to analysis and plotting in Python notebooks:

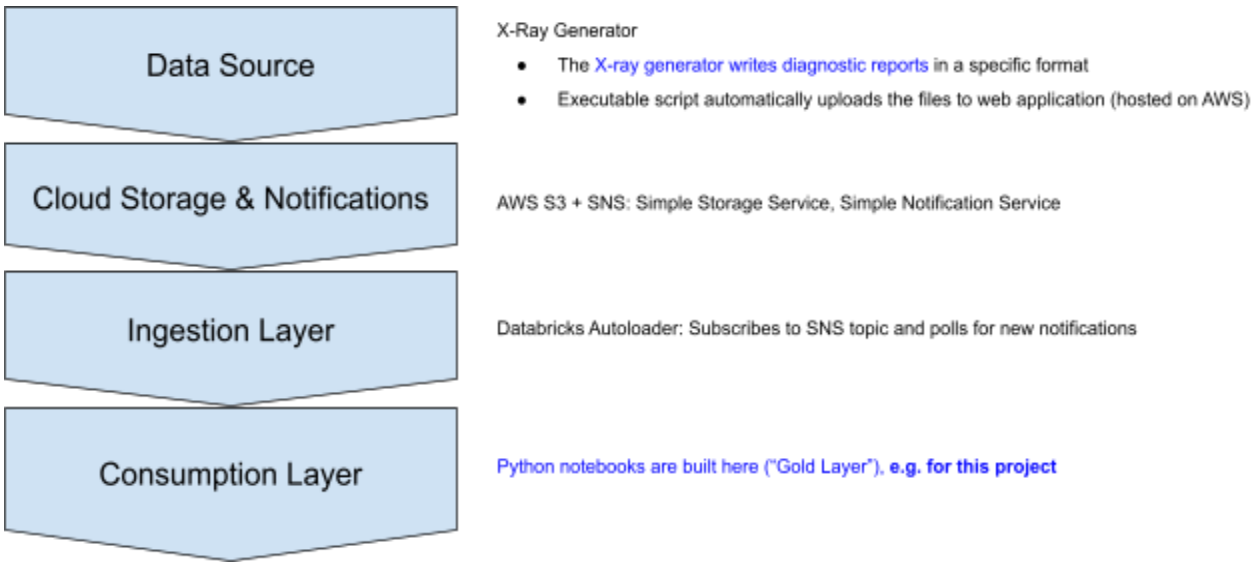


Figure 3: Data flow diagram

The Consumption layer, where work for this project will take place, corresponds to the Gold layer according to the Lakehouse architecture.

7 Data Model

The data model at the conceptual level, the logical level and the physical level is explained just briefly, the data model design is not part of this project, since the data is already available on Databricks consumption layer, ready to work with.

Conceptual Data Model

On the conceptual level, there are entries per row in the following form:

- Timestamp (Date + Time)
- Device ID (Generator)
- Event type (warm-up, workpoint change, arc, exposure, ...)
- Operating point, consisting of high voltage in kV, emission current in mA, ...
- Temperature, measured at several locations in the generator device
- Focal spot
- And more

Logical Data Model

The names of the 36 columns are (raw data table):

```
['time', 'device', 'type', 'voltage', 'current', 'filament current',
'grid voltage', 'power', 'temperatures', 'focal spot', 'focal spot max power',
'mid exposure focal spot change', 'length', 'duration', 'components',
'shutdown type', 'reason', 'tube', 'component change aTank',
'component change cTank', 'component change Software', 'component change Tube',
'component change IFC', 'component change ECU', 'component change POC1',
'component change POC2', 'temp CTank', 'temp ATank', 'temp ECU-Oil',
'temp POC2', 'temp FGU-Oil', 'temp POC1', 'temp IFC', 'cable length',
'created at', 'timestamp']
```

Physical Data Model

Infrastructure used is Amazon Web Services AWS cloud storage for uploading diagnostic reports from devices, and data ingested in Databricks where the compute clusters for analysis, statistical calculation, model training etc. are.

Raw data table size in Databricks:

- 709 MiB (mebibyte: 1 MiB = 2^{20} bytes = 1,048,576 bytes)
- rows x columns: 25’762’956 x 36, this number of rows is the current one and is growing over time with new entries from the field devices

8 Documentation

The project will be documented in the form of Python notebooks, consisting of code and comments. Depending on which sub-topic (statistical test, regression, clustering) additionally in slides and/or a report, since they probably will be the Module 2, 3, .. or Final Project work.

9 Risks

What can go wrong is that arc event data may be biased resp. too pessimistic, as mentioned in “5 Data Quality”. Regression and clustering tasks may have such a bad performance that nothing useful comes out (degradation can not be predicted with sufficient confidence, no clear, understandable patterns found, ...).

Regarding the arc topic, the extreme values or outlier devices could be analyzed in detail, since sometimes specific devices, e.g. running in our internal R&D lab, face extraordinary stress.

Results from biased data can be misleading of course, and where possible it will be compensated, where the time consumed for that will be missing at another place.

10 Preliminary Studies

First results for answering the question “Is the accumulated number of tube arcs (type of event) per accumulated exposure time (in hours) per generator device below a certain threshold, e.g. 4’000 arcs per 10’000 hours* (rate 0.4 arcs per hour)?” are shown.

Descriptive statistics for arcs per generator device and exposure time per generator device:

- Total number of arcs: 54’626
- Summary of number of arcs per device statistics:
    - min: 0

    - max: 11894
    - mean: 68.28
    - median: 4.0
    - std: 468.496
- Total exposure time: 780’570 hours
- Summary of exposure time per device statistics:
    - min: 0.000167
    - max: 25735.39
    - mean: 1015.046
    - median: 181.878
    - std: 2611.528
- No normal distribution: Skewness/overdispersion (from arcs and exposure time per device statistics)

The following plot gives an overview regarding arcs per exposure per device. Each point represents a generator device. Tube type groups are plotted in different colos, to check for related pattern (e.g. 160kV seems to be less problematic than 225kV). The red line drawn from (0, 0) to (10’000 h, 4000 arcs) splits the devices into “good” and “bad”ones.

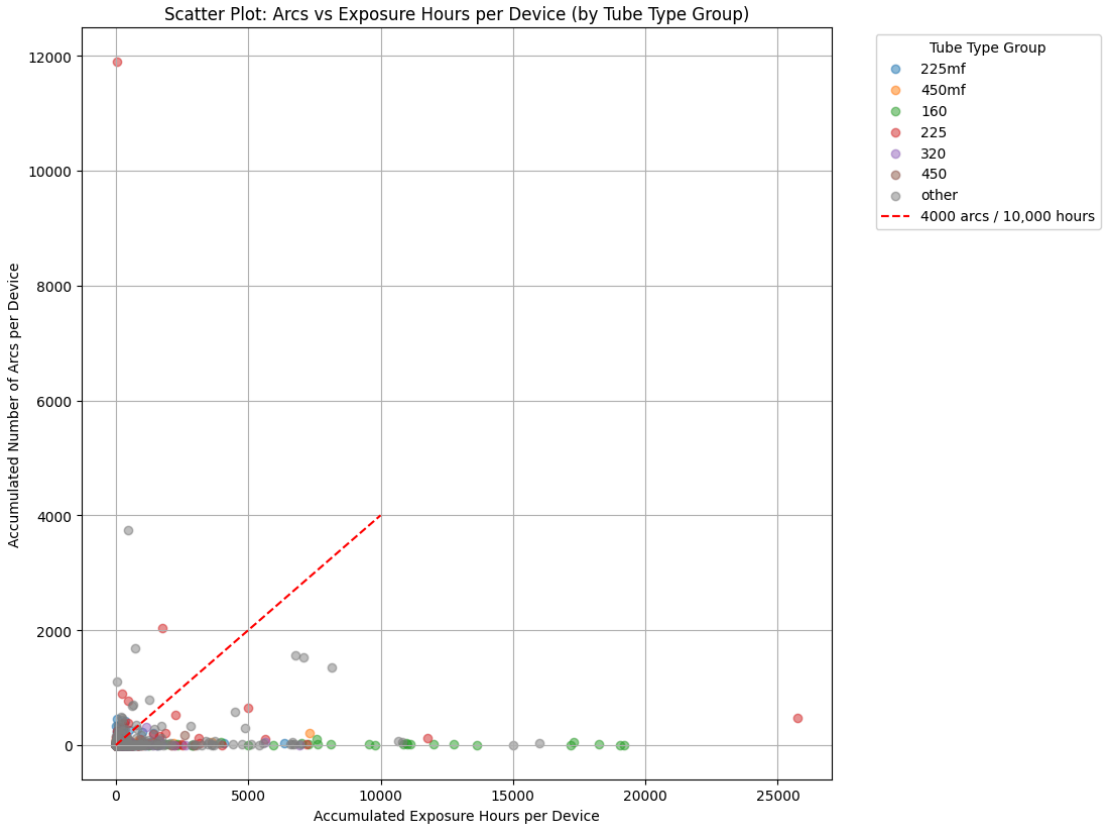


Figure 4: Scatter plot of number of arcs per exposure time per generator device.

11 Conclusions

For the statistical test answering the question whether the number of arcs per exposure time and generator device is below a certain threshold, beside the descriptive statistics shown under “10 Preliminary Studies”, there is already a result for the statistical test existing, each for fleet-level and device-level separately. The fake arcs topic introduced in “9 Risks” is not handled yet, if needed it can be implemented quite straight-forward because it was done in the past for another use case.

Since the field data is available in Databricks, it could be started immediately with the pre-processing and model training for the regression, clustering and anomaly detection tasks. It has to be decided with which to start with, depending on priority and estimated effort.

For the supervised learning case, where a degradation of a component or subsystem is tried to be predicted, first analysis shows the effect for a subset of tubes. It is to be discovered, which input variable has which contribution on degradation: Age (time since manufacturing resp. first commissioning), exposure time, or number of cycles.

Regarding the unsupervised learning topic, where patterns and anomalies shall be discovered, the suspicious devices found in the arcs per exposure time per device can be used for plausibility checks for anomalies. For patterns one could start which continuous vs cycling mode operation and compare to which tube type is mainly used per group, since there are ones optimized for continuous mode, and others for cycling mode.

Statement

The following part is mandatory and must be signed by the author or authors.

„Ich erkläre hiermit, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen benutzt habe. Alle Stellen, die wörtlich oder sinngemäss aus Quellen entnommen wurden, habe ich als solche gekennzeichnet. Mir ist bekannt, dass andernfalls die Arbeit als nicht erfüllt bewertet wird und dass die Universitätsleitung bzw. der Senat zum Entzug des aufgrund dieser Arbeit verliehenen Abschlusses bzw. Titels berechtigt ist. Für die Zwecke der Begutachtung und der Überprüfung der Einhaltung der Selbstständigkeitserklärung bzw. der Reglemente betreffend Plagiate erteile ich der Universität Bern das Recht, die dazu

erforderlichen Personendaten zu bearbeiten und Nutzungshandlungen vorzunehmen, insbesondere die schriftliche Arbeit zu vervielfältigen und dauerhaft in einer Datenbank zu speichern sowie diese zur Überprüfung von Arbeiten Dritter zu verwenden oder hierzu zur Verfügung zu stellen."

Date: Signature(s):

2025-10-05

References and Bibliography

[1] https://www.databricks.com/product/data-intelligence-platform

[2] https://www.databricks.com/spark/about

[3] https://spark.apache.org/docs/latest/api/python/index.html

[4] https://docs.databricks.com/aws/en/lakehouse/medallion