PYCON 2013
SANTA CLARA, CA • MARCH 13-21

SAFRAN
Morpho

Data Classification using Python and R – Industry Use

Meenal Pant
Morpho Detection (MDI), 7151 Gateway Blvd, Newark, CA 94560

Introduction

1. Why is "data classification" needed ?
Today the world is data driven world … data classification and data mining is critical to understanding patterns and behavior. This enables data prediction and projection.

2. What is data classification ? … and Data mining ?
Data mining is typically done on abstract data think social networking data**. Data Classification** is typically done on known data.

At MDI, we manufacture complex, state of the art baggage scanning equipment. The many components of these EDS's (Explosive Detection System) produce a large amount of known data. We ***classify, process & predict*** this data to provide proactive maintenance

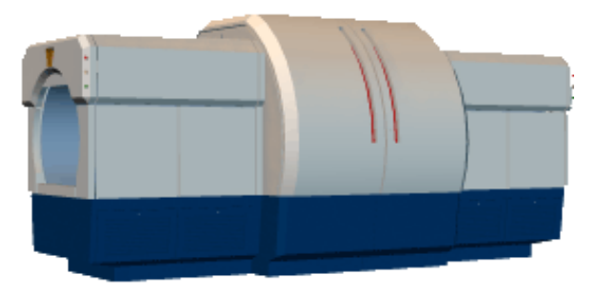


Figure 1. Explosive Detection System (EDS)

Process

- Collect data from the airport(s)
- Analyze large datasets daily and publish on a web dashboard
- Define critical parameters
- Use rules engine and knowledge base to create alerts
- Email alerts to relevant staff

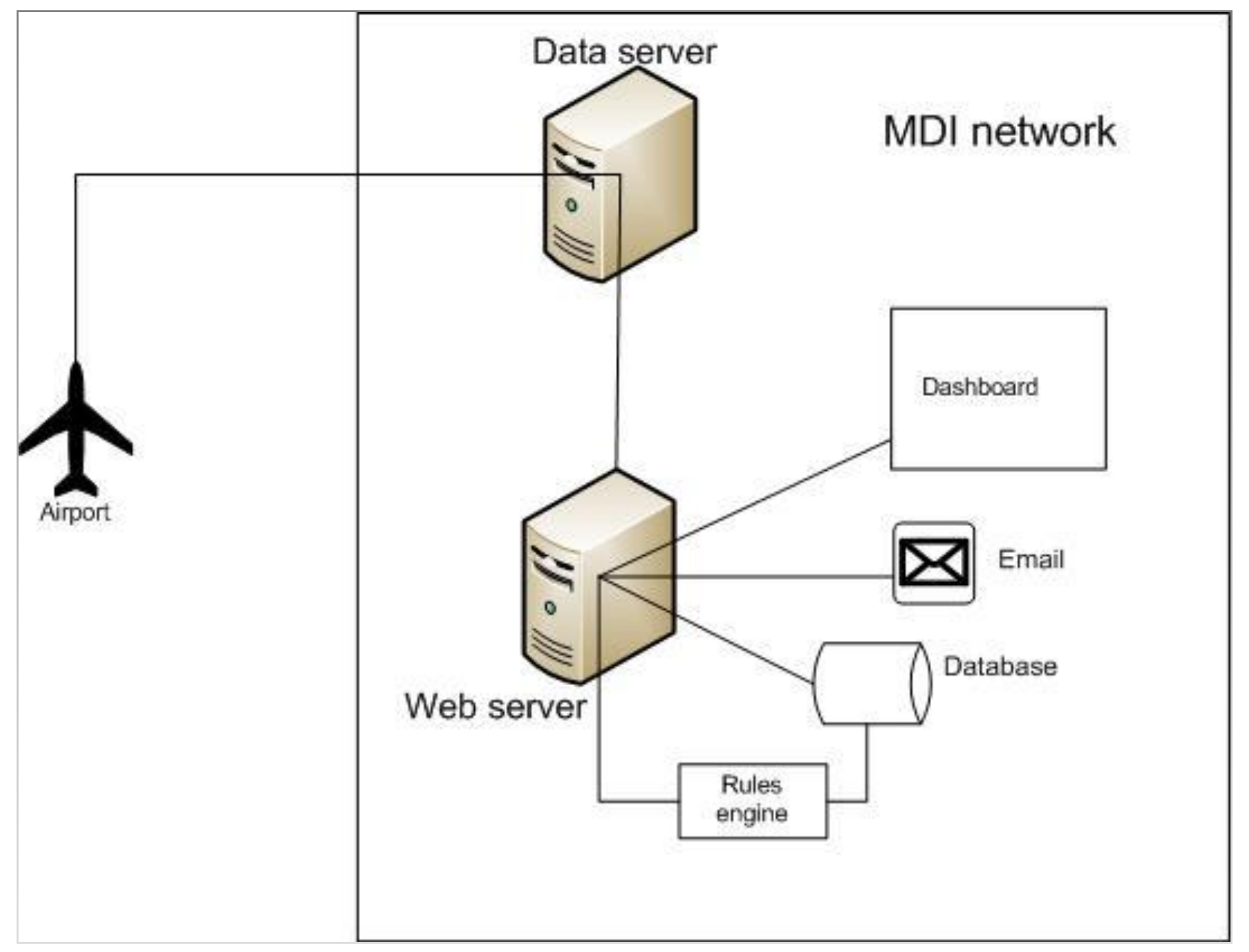


Figure 2. Top level design for the real time performance dashboard.

Technology

python powered   django   R   RPy

Decision Tree Learning
From Wikipedia, Decision Tree Learning used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value.
In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making.
http://en.wikipedia.org/wiki/Decision_tree_learning

Classification And Regression Tree (CART) model is used for "rules engine" here.

```
gData <-read.csv("enc1err.csv", sep=",", header=TRUE)
fit <- rpart(Result ~ err  + errf, method="class", data=gData)
```

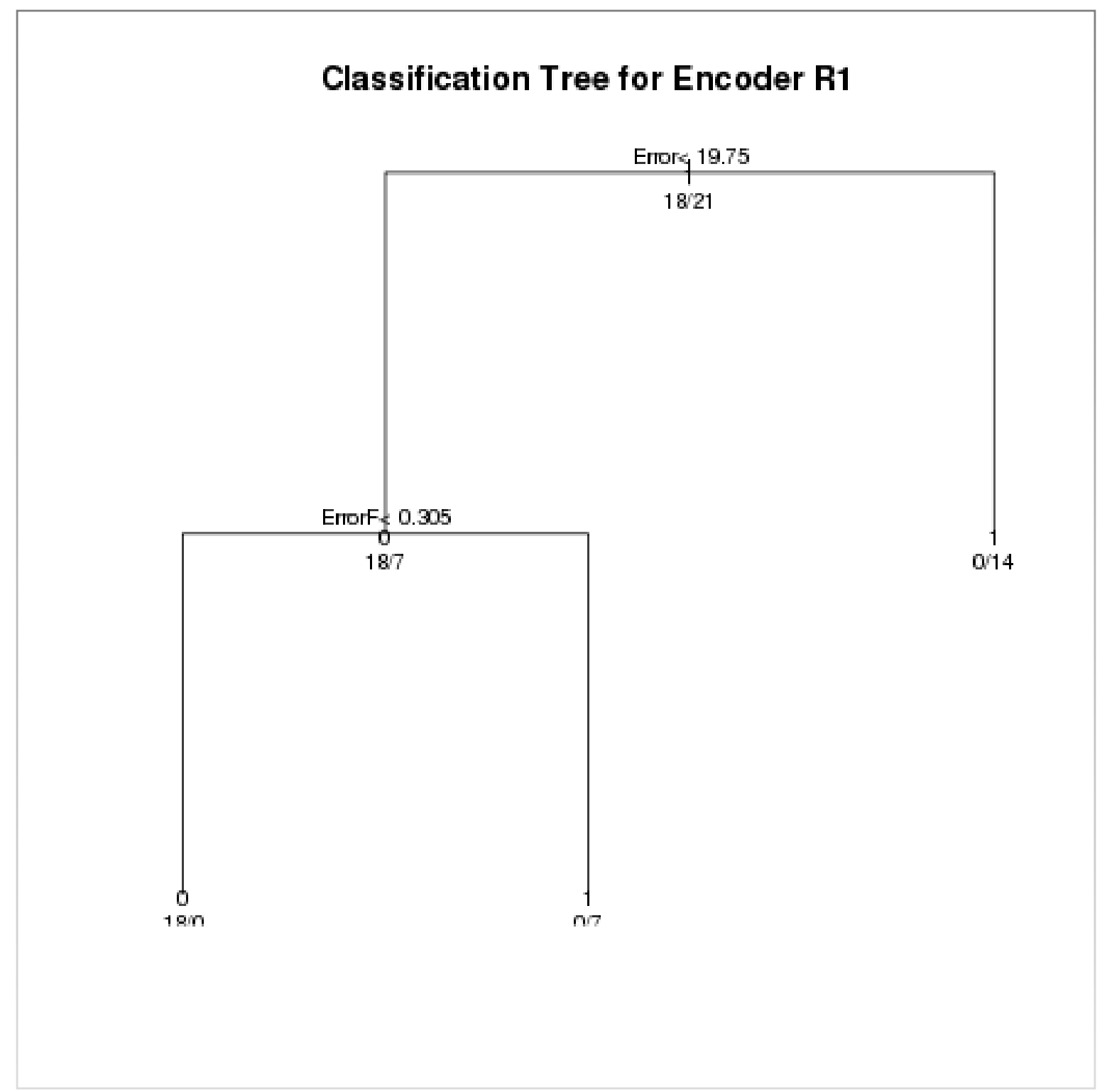


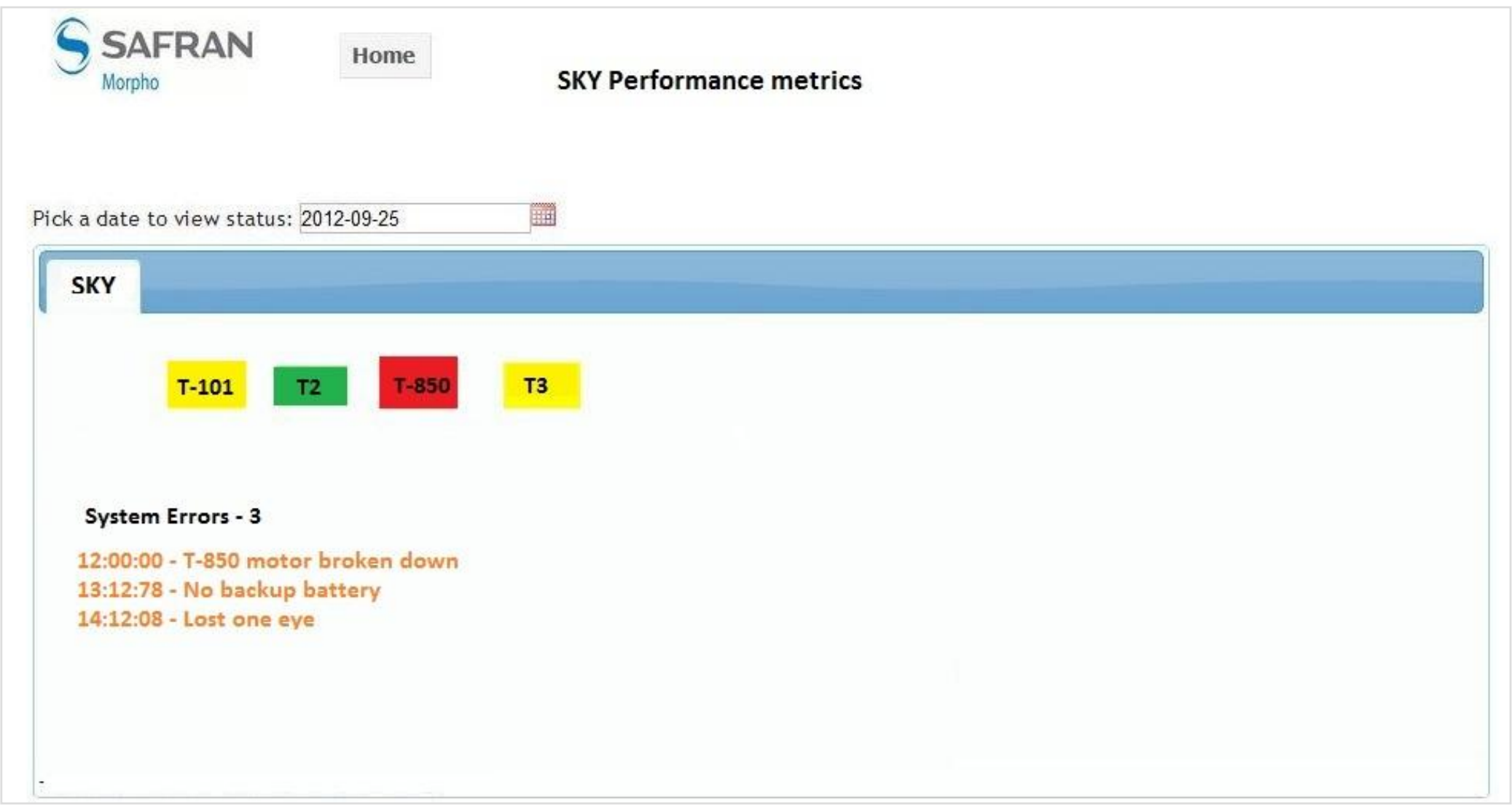
Figure 3. Decision Tree Classification for one of the key components in the MDI EDS.

R is a free software package for statistical computing and graphics. rpart is the package for producing CART models
To create an R decision tree:

- Get training samples
- 'Fit' data using rpart
- Score using validation data
- Predict test data
- Additional steps, pruning a tree etc.

Implementation

Color codes: Yellow = warning, Red = Error , Green = Normal



SKY Performance metrics

Pick a date to view status: 2012-09-25

SKY

T-101   T2   T-850   T3

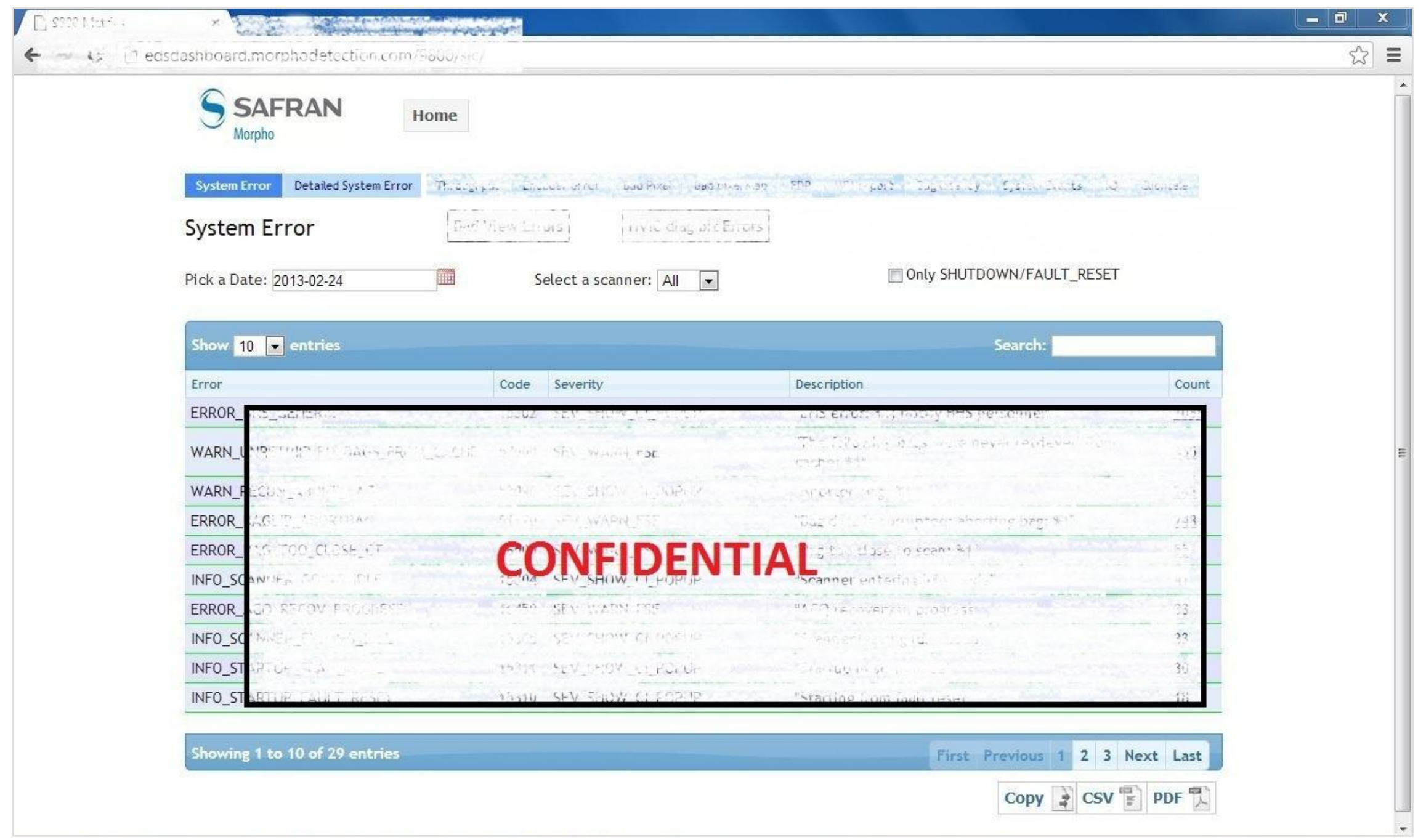
System Errors - 3
12:00:00 - T-850 motor broken down
13:12:78 - No backup battery
14:12:08 - Lost one eye

Figure 4. Example of what a status page looks like on the web based dashboard.

Sample Code

```
#rpy script to predict and score
import os
from config import ROOT_PATH
from rpy2 import robjects
from rpy2.robjects import r
rpart = r.library("rpart")
class Rules(object):
    #Create rules based on CART model
    def __init__(self):
        self.csvpath = os.path.join(ROOT_PATH, "/train/")
    def r1tree(self):
        #Fitting the tree based on training sample
        self.e1path = os.path.join(self.csvpath,
"e1err.csv")
        self.r1cmd = 'e1Data <-read.csv("%s", sep=",",
header=TRUE)'%(self.e1path)
        robjects.r(self.r1cmd)
        robjects.r('e1fit <- rpart(res ~ err + ef,
method="class", data=e1Data)')
     def r1val(self):
        #Validate and predict new data
        robjects.r('validationData <-
read.csv("/tmp/e1val.csv", sep=",",   header=TRUE)')
        robjects.r('test <- predict(e1fit,interval =
"prediction",newdata=validationData, type="vector")')
        self.scored = robjects.r('test')
        return self.scored
```

✓ Returns a score of 1 or 0 and classifies the data
✓ Data analysis can now be performed by applying rules on this classified data.



System Error

Pick a Date: 2013-02-24   Select a scanner: All   □ Only SHUTDOWN/FAULT_RESET

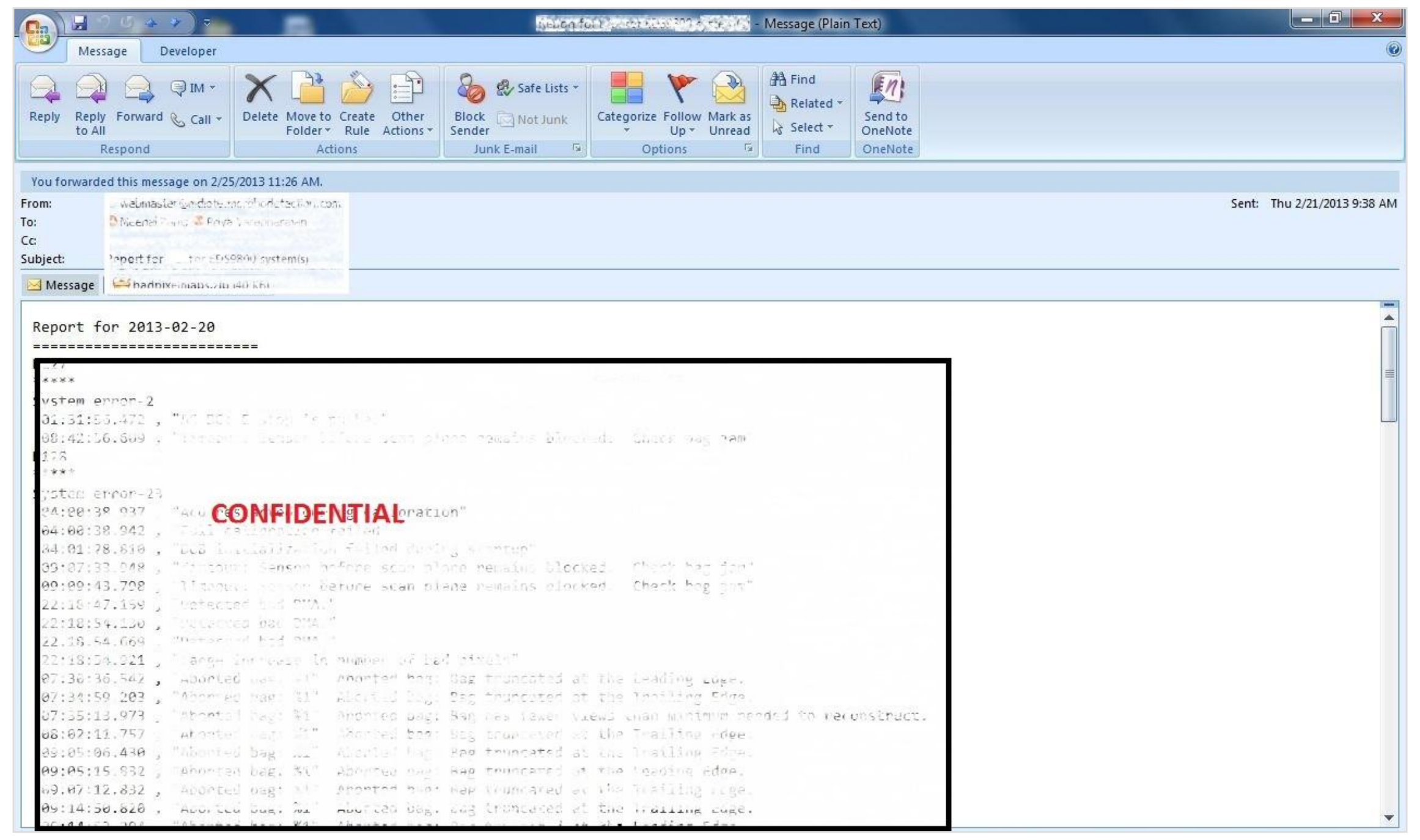
Show 10 entries   Search:

Error	Code	Severity	Description	Count
ERROR_	[illegible]	[illegible]	[illegible]	[illegible]
WARN_U	[illegible]	[illegible]	[illegible]	[illegible]
WARN_R	[illegible]	[illegible]	[illegible]	[illegible]
ERROR_	[illegible]	[illegible]	[illegible]	[illegible]
ERROR_	[illegible]	[illegible]	[illegible]	[illegible]
INFO_SC	[illegible]	[illegible]	[illegible]	[illegible]
ERROR_	[illegible]	[illegible]	[illegible]	[illegible]
INFO_SC	[illegible]	[illegible]	[illegible]	[illegible]
INFO_ST	[illegible]	[illegible]	[illegible]	[illegible]
INFO_ST	[illegible]	[illegible]	[illegible]	[illegible]

CONFIDENTIAL

Showing 1 to 10 of 29 entries   First Previous 1 2 3 Next Last

Figure 5. Detailed reports example



You forwarded this message on 2/25/2013 11:26 AM.

From:
To:
Cc:
Subject:
Sent: Thu 2/21/2013 9:38 AM

```
Report for 2013-02-20
=========================
```

CONFIDENTIAL

Figure 6. Daily email to the maintenance support staff

Benefits

- Direct cost benefits due to reduction in downtime
- Can identify and troubleshoot issues faster
- Improves personnel efficiency
- Always available status and results
- Archived results to learn trends and patterns

Acknowledgments

Thanks to Samit Basu, Chief Engineer MDI , and Natarajan Subramanian, Senior Software Manager MDI, for their guidance with this research.

Thanks to MDI and Safran for sponsoring and enabling me to present my research at PyCon 2013.

Questions and feedback

Please contact *mpant@morphodetection.com*