Metadata Management - Code Demonstration

Karthikeyan Chokappa (KC)

2021-04-23

Contents

Problem Statement

Demonstrate code to monitor the Metadata Management Service Level Objective (SLO) related to Data Governance.

Assumption

Data Governance score for each field or parameter is calculated based on various criteria for each data types. Data goodness is represented between 0.001 to 0.999, where 0.001 is low quality and 0.999 is high quality data, whereas values 0 and 1 represent anomalous data. For some typical data types the data goodness is estimated as follows,

- Numeric - Percentage of good data values is estimated by marking "NULL" or Outliers (Estimated using Box-plot method from the last 7 days) as bad data
- String - Percentage of good data values is estimated by marking "NULL" or values with highest proportion of special characters as bad data

For the purpose of this use case demonstration the goodness of data for 10 different fields or parameters for a given input table (table_01) for 30 days is simulated and anomalies are injected randomly. The 10 fields represent the information needed to complete a transaction and each filed is assumed to have equal importance.

Metadata Governance Monitoring Method

Pipeline Code Demonstration

Load Packages / Libraries

```
library(qicharts2)
```

Define Constants

```
pp_wrkdir <- getwd()
```

Load Simulated Governance Data (with Failures Injected)

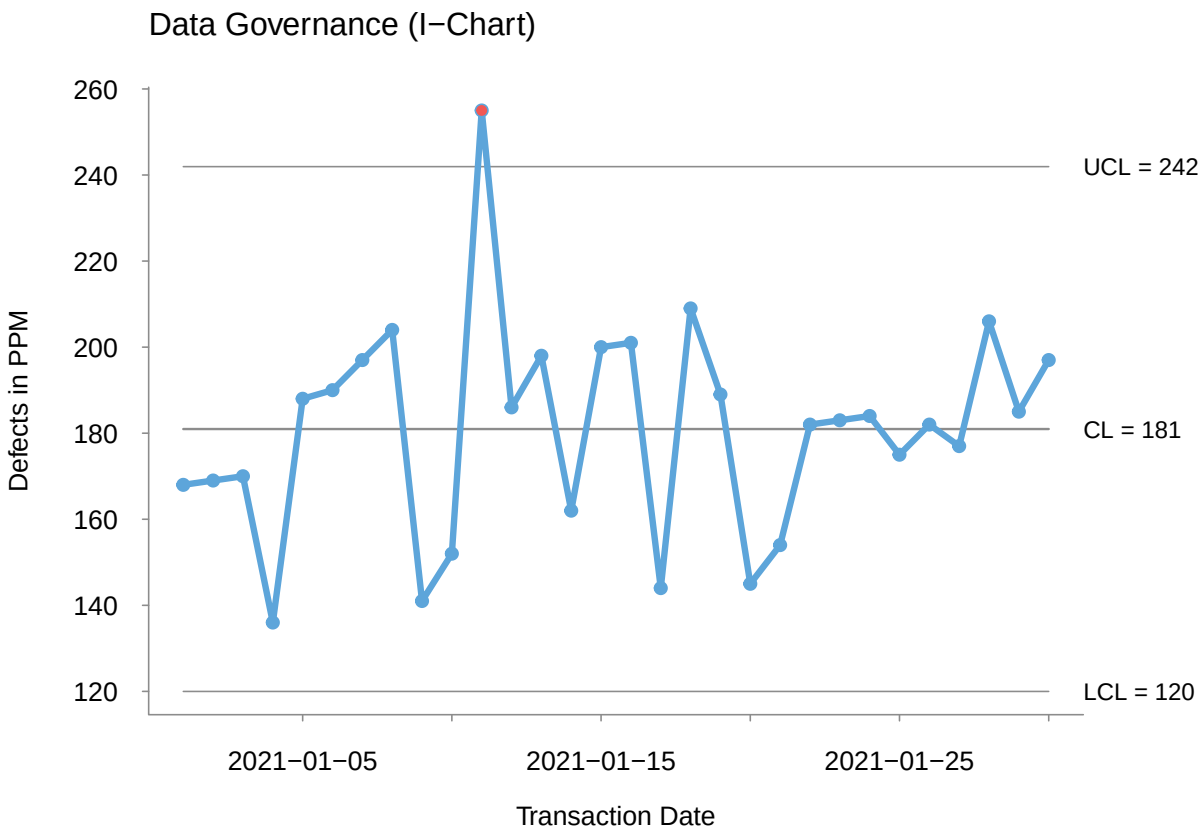
```
#Simulation - Parameters
Transactions_Count            <- 1000
Transactions_DefectsInPPM_Avg <- 178
Transactions_DefectsInPPM_Std <- 19

#Simulation - Data
Transactions <- rpois(30, lambda = Transactions_Count)
DefectsInPPM <- round(rnorm(30,
                            mean = Transactions_DefectsInPPM_Avg,
                            sd   = Transactions_DefectsInPPM_Std))
Date         <- seq(as.Date('2021-01-01'),length.out = 30, by = 'day')
dbGovernance <- data.frame(Date, Transactions, DefectsInPPM)

#Simulation - Failures Injected
dbGovernance$DefectsInPPM[11] <- 255
```

Monitor Governance Quality using Individual Control Chart

```
# Plot I-Chart of Governance
qicharts::qic(y = dbGovernance$DefectsInPPM,
              x = dbGovernance$Date,
               chart = 'i',
               main  = 'Data Governance (I-Chart)',
    ylab  = 'Defects in PPM',
    xlab  = 'Transaction Date')
```



Summary

The above is just sample representation demonstrating a typical monitoring mechanism for Metadata Governance information over daily transactions. This information can be pulled for all parameters for all databases across any given application. Using anomaly or defect aggregation method the overall health score of data platforms can be monitored and alerts sent if the health score / index degrades significantly.