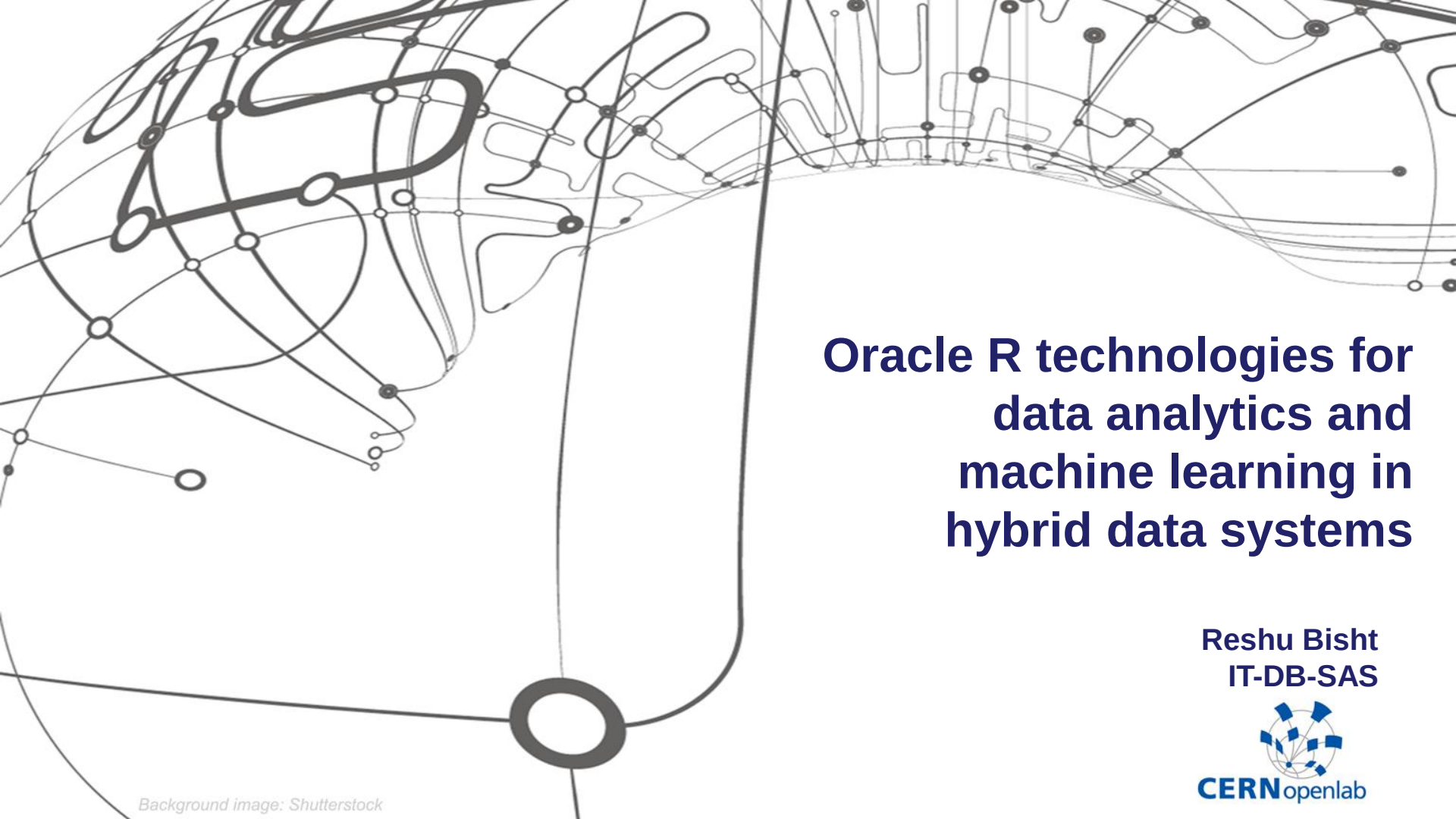


Oracle R technologies for data analytics and machine learning in hybrid data systems

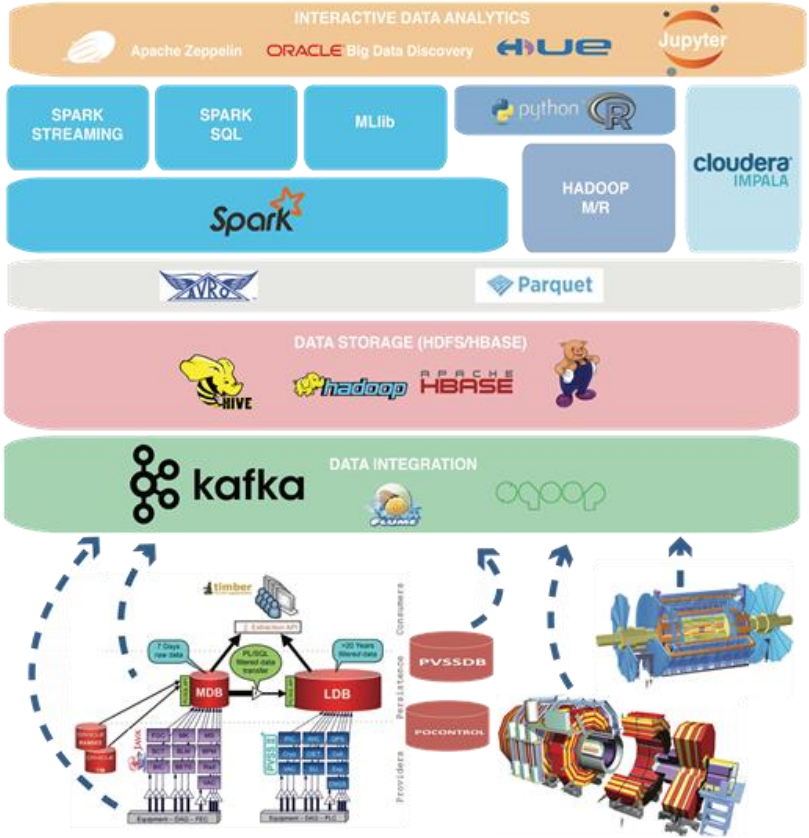
Reshu Bisht
IT-DB-SAS

CERN openlab

Background image: Shutterstock

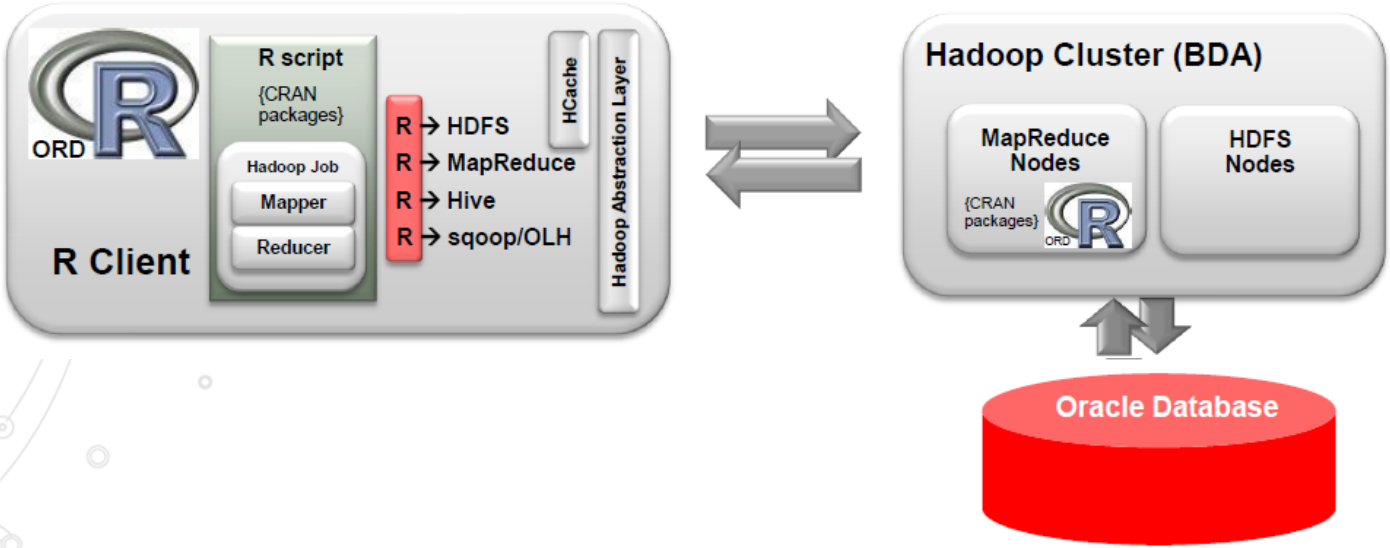
Introduction

- A wide variety of CERN's uses cases require efficient analysis.
- The Volume, Velocity and Variety of Data in CERN
- New scalable data services available
- Evaluating Scalable Analytic Solutions on hybrid data platforms



Background image: Shutterstock

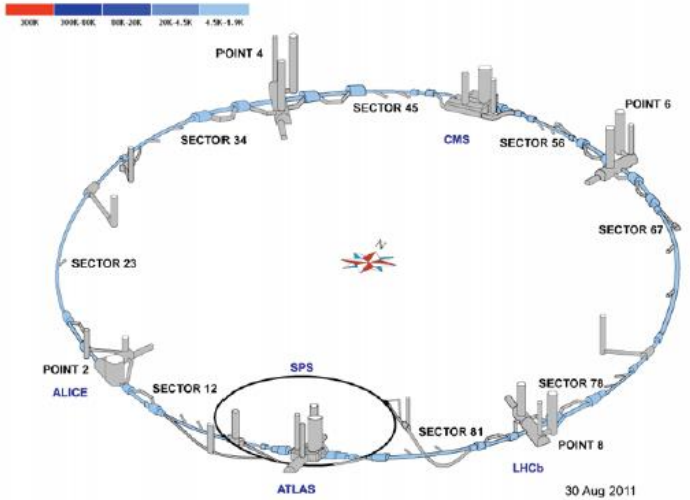
Oracle R Advanced Analytics for Hadoop



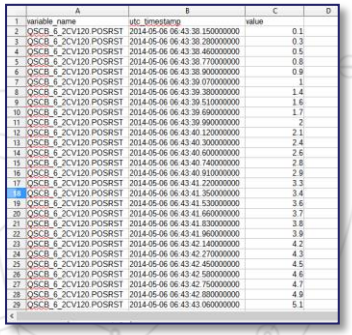
Use Case 1: Cryo Valves Degradation Analysis

- One of the largest Cryogenic Installations
- Data from QSCB_6_2CV120
- Time Series data set
  - Timestamp
  - Order
  - Feedback
  - Error : Order - Feedback

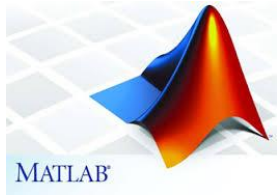
Instrument/Actuators	Total
Temperature [1.6 – 300 K]	10361
Pressure [0 – 20 bar]	2300
Level	923
Flow	2633
Control valves	3692
On/Off valves	1835
Manual valves	1916
Virtual flow meters	325
Controllers (PID)	4833

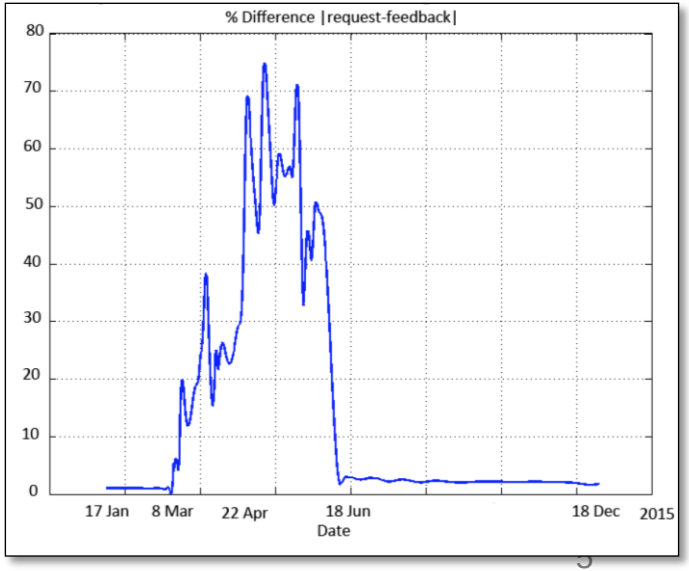


- **Traditional data analysis workflow has scalability issues**
  - **Extract/move data is very expensive**
  - **Local processing non powerful machines**
  - **Difficult to process in a distributed manner**



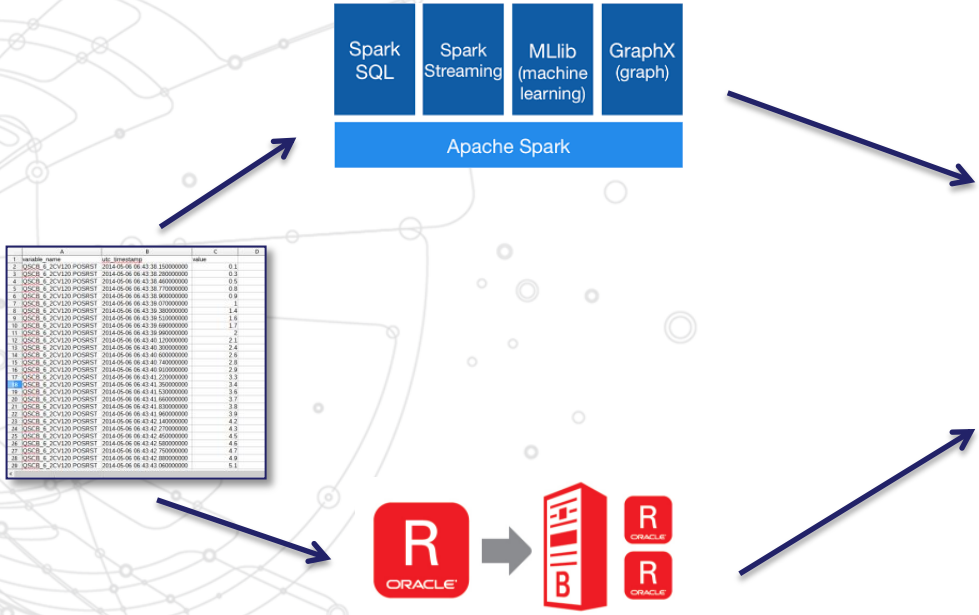
	A	B	C	D
1	variable_name	utc_timestamp	value	
2	QSCB_6_2CV120.POSRST	2014-05-06 06:43:38.150000000	0.1	
3	QSCB_6_2CV120.POSRST	2014-05-06 06:43:38.280000000	0.3	
4	QSCB_6_2CV120.POSRST	2014-05-06 06:43:38.460000000	0.5	
5	QSCB_6_2CV120.POSRST	2014-05-06 06:43:38.770000000	0.8	
6	QSCB_6_2CV120.POSRST	2014-05-06 06:43:38.900000000	0.9	
7	QSCB_6_2CV120.POSRST	2014-05-06 06:43:39.070000000	1	
8	QSCB_6_2CV120.POSRST	2014-05-06 06:43:39.380000000	1.4	
9	QSCB_6_2CV120.POSRST	2014-05-06 06:43:39.510000000	1.6	
10	QSCB_6_2CV120.POSRST	2014-05-06 06:43:39.690000000	1.7	
11	QSCB_6_2CV120.POSRST	2014-05-06 06:43:39.990000000	2	
12	QSCB_6_2CV120.POSRST	2014-05-06 06:43:40.120000000	2.1	
13	QSCB_6_2CV120.POSRST	2014-05-06 06:43:40.300000000	2.4	
14	QSCB_6_2CV120.POSRST	2014-05-06 06:43:40.600000000	2.6	
15	QSCB_6_2CV120.POSRST	2014-05-06 06:43:40.740000000	2.8	
16	QSCB_6_2CV120.POSRST	2014-05-06 06:43:40.910000000	2.9	
17	QSCB_6_2CV120.POSRST	2014-05-06 06:43:41.220000000	3.3	
18	QSCB_6_2CV120.POSRST	2014-05-06 06:43:41.350000000	3.4	
19	QSCB_6_2CV120.POSRST	2014-05-06 06:43:41.530000000	3.6	
20	QSCB_6_2CV120.POSRST	2014-05-06 06:43:41.660000000	3.7	
21	QSCB_6_2CV120.POSRST	2014-05-06 06:43:41.830000000	3.8	
22	QSCB_6_2CV120.POSRST	2014-05-06 06:43:41.960000000	3.9	
23	QSCB_6_2CV120.POSRST	2014-05-06 06:43:42.140000000	4.2	
24	QSCB_6_2CV120.POSRST	2014-05-06 06:43:42.270000000	4.3	
25	QSCB_6_2CV120.POSRST	2014-05-06 06:43:42.450000000	4.5	
26	QSCB_6_2CV120.POSRST	2014-05-06 06:43:42.580000000	4.6	
27	QSCB_6_2CV120.POSRST	2014-05-06 06:43:42.750000000	4.7	
28	QSCB_6_2CV120.POSRST	2014-05-06 06:43:42.880000000	4.9	
29	QSCB_6_2CV120.POSRST	2014-05-06 06:43:43.060000000	5.1	

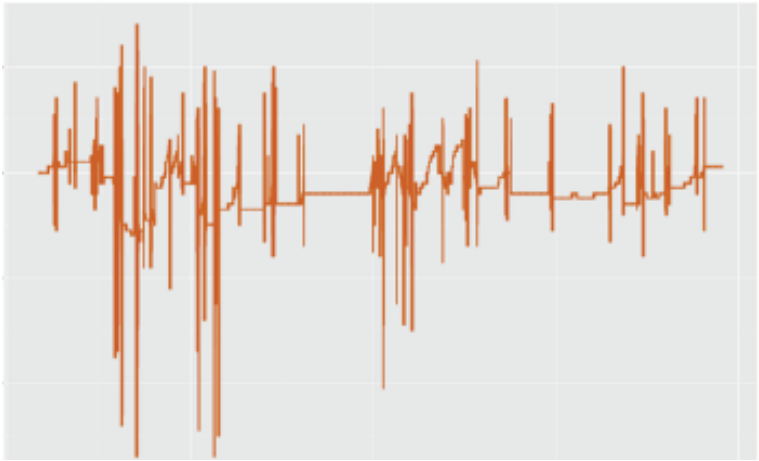




Background image: Shutterstock

Implementation in Spark and ORAAH





Use Case 2: Pattern Classification for Faulty Cryo Valves

CV910
CV943
CV947

Pre-Processing

S= Order- Feedback

Result

Training

Testing

Mahalanobis Distance	0.94
Support Vector Machine	0.96

Feature Set:

- Max Error
- Variance
- Noise Band: B(S)
- Rope Distance: R(S)

Scalable Platform

ORAAH

```
library(ORCH)
spark.connect(master="yarn-client",memory="2G",
              dfs.namenode="bigdatalite.localdomain")
train  <- hdfs.put(training)
model  <- orch.ml.svm(formula = status ~ rope_dist +
                      bs + mean + var + max ,data = train)
test <- hdfs.put(testing)
pred   <- predict(model, newdata = test)
hdfs.write(pred, outPath = "Prediction")
```

Spark

```
import org.apache.spark.mllib.classification.{SVMModel, SVMWithSGD}
import org.apache.spark.mllib.evaluation.BinaryClassificationMetrics
import org.apache.spark.mllib.util.MLUtils
import org.apache.spark.SparkContext._

val train = MLUtils.loadLibSVMFile(sc, "/user/r1/rtr.data")
val test = MLUtils.loadLibSVMFile(sc, "/user/r2/rte.data")

val numIterations = 90
val model = SVMWithSGD.train(train, numIterations)

model.clearThreshold()

val scoreAndLabels = test.map { point =>
  val score = model.predict(point.features)
  (score, point.label)
}

val metrics = new BinaryClassificationMetrics(scoreAndLabels)
val auROC = metrics.areaUnderROC()

println("Area under ROC = " + auROC)

model.save(sc, "/home/cloudera/d")
val sameModel = SVMModel.load(sc, "target
                              /tmp/scalaSVMWithSGDModel")
```

Future Work

- **Scale the analysis in bigger Hadoop cluster using larger cryo datasets**
- **Test & Benchmark with public datasets**
- **Give feedback to Oracle about performance and features once we have final results**
  - Improving the product: Two issues were found
    - BigDataLite VM – Ground Configuration
    - Orch package – sampling error

Conclusion

- **R ecosystem is good for data analysis**
- **ORAAH**
  - No pains in adapting code from R
- **Hybrid Data Systems looks quite promising to cover multiple scenarios**

Thank You

Reshu Bisht

Supervisors: Manuel Martin Marquez
Antonio Romero Marín
Mark Hornick (Oracle)