

# Integrating HP Vertica with External Analytics Engines:

A case for Spark and Distributed R

Jeff LeFevre, Rui Liu, Malu Castellanos, Qiming Chen, Meichun Hsu

HP Vertica West  
XLDB 2015

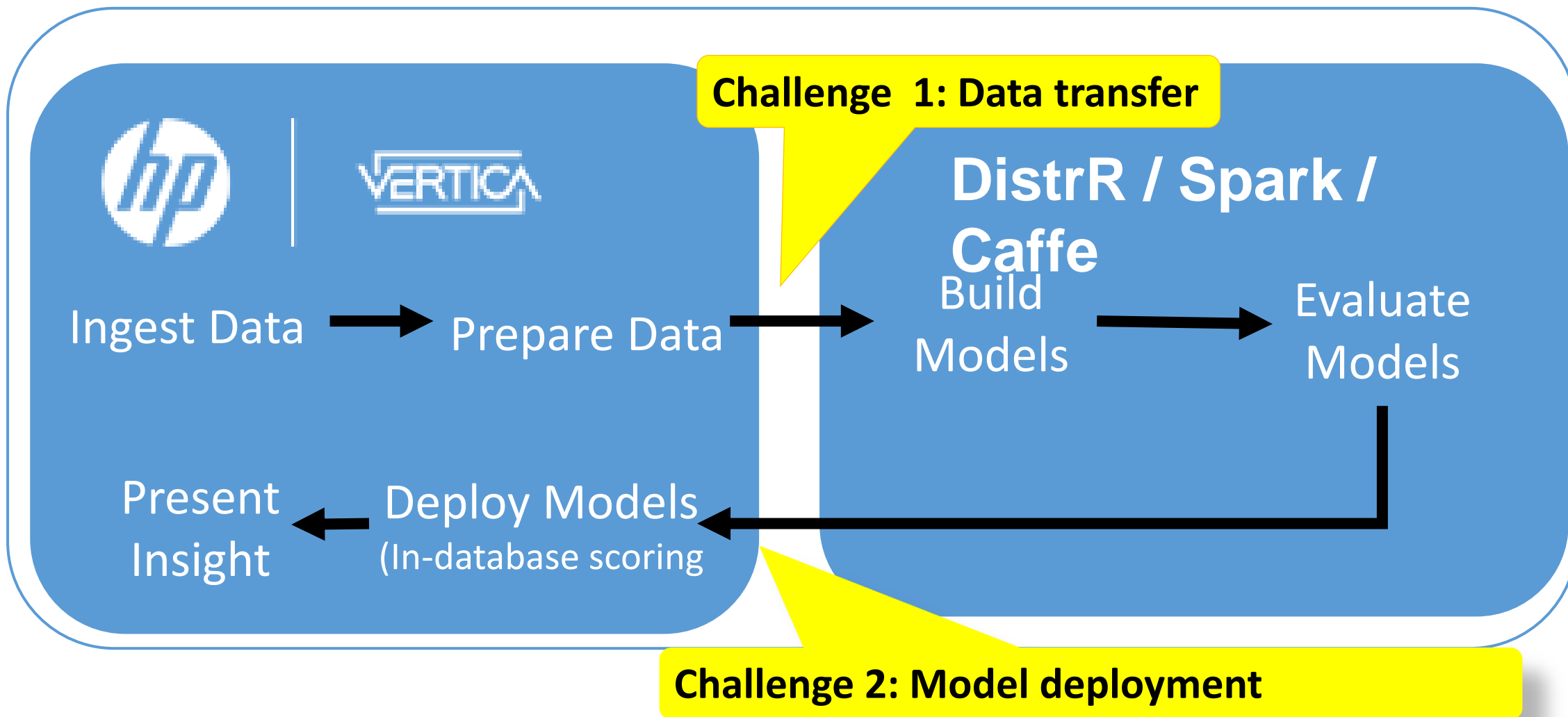
# High Level Goals

- Extend Vertica's capability with additional ML algorithms
- Keep database as the enterprise quality data manager
- Let data scientists use tools they are familiar with
- Easily deploy ML models within the database

# Three Current Efforts at HP Vertica

- Integration with Distributed R (HP Labs project)
  - Distributed ML platform with several ML algorithms
- Integration with Spark
  - Distributed ML platform with MLlib, GraphX
- Integration with Caffe
  - Deep learning platform

# HP Vertica + DistR / Spark / Caffe



# Example Distributed R User Session

## # LOAD DATA

```
$: data <- db2darray(TABLE1, list('def'), list('A', 'B'))
```

## # BUILD MODEL

```
$: model <- hpdglm(data$Y, data$X, family=binomial, ...)
```

## # DEPLOY MODEL

```
$: deploy.model(model, 'my_model')
```

## # IN-DATABASE PREDICTION

```
$: query <- SELECT glmPredict(A, B,  
                             using PARAMS model='my_model')  
           from TABLE2;
```

```
$: res <- sqlQuery(conn, query)
```

# Example Spark User Session

## # LOAD DATA

```
$: val data = new VerticaRDD(sc, connection=getConnection,  
table, numPartitions=numOfPart, mapRow=extractValues)
```

## # BUILD MODEL

```
$: val clusters = KMeans.train(data, numClusters, numIterations)
```

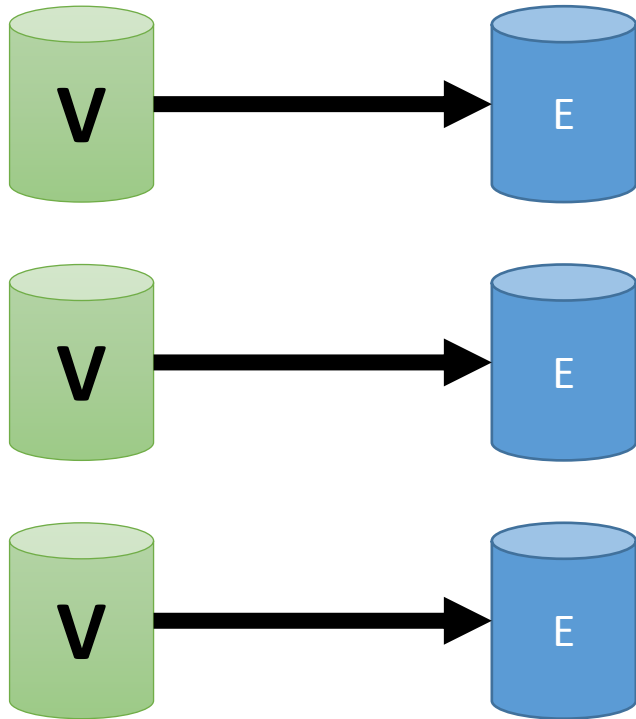
## # DEPLOY MODEL

```
$: val vert = new ModelSerDes(connection=getConnection, table)  
$: vert.store(clusters)
```

## # IN-DATABASE PREDICTION

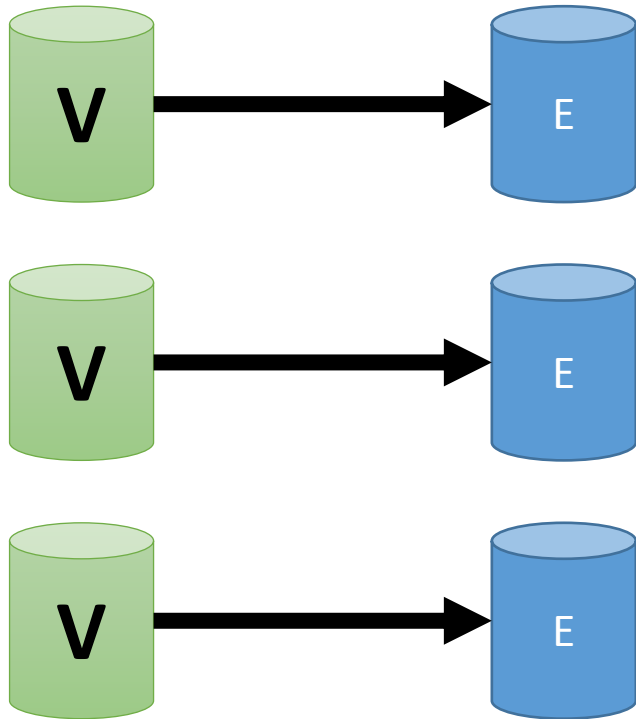
```
$: select spark_km_predict(x,y,z USING PARAMETERS modelId=2066188038)  
from table;
```

# Data Transfer Issues

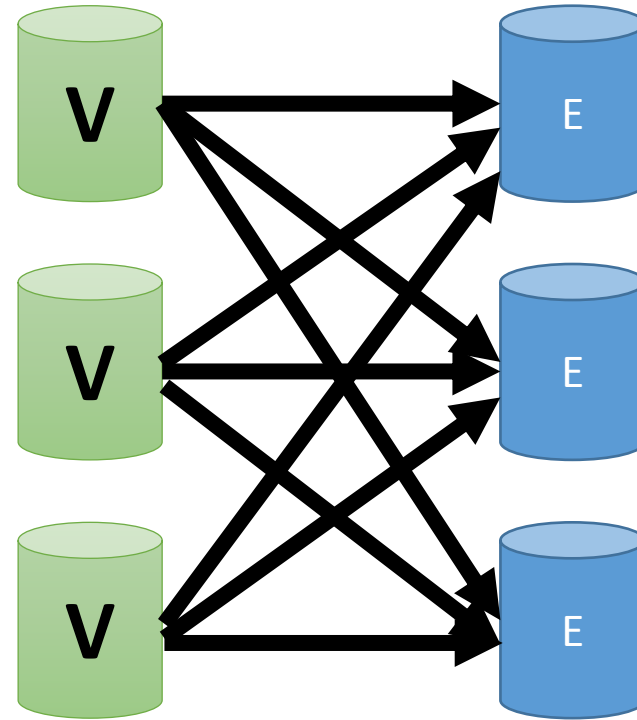


Locality Preserving Policy

# Data Transfer Issues



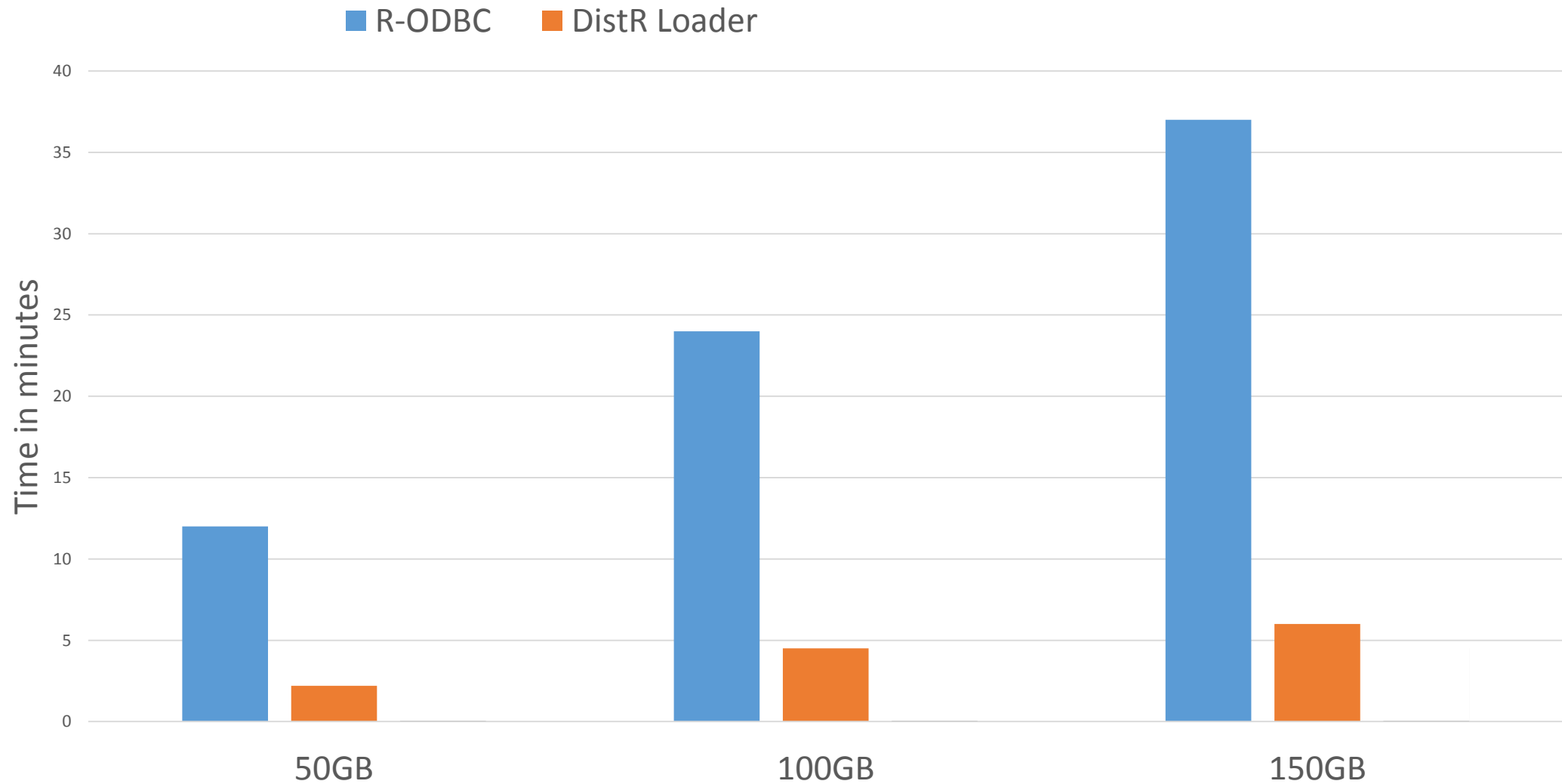
Locality Preserving Policy



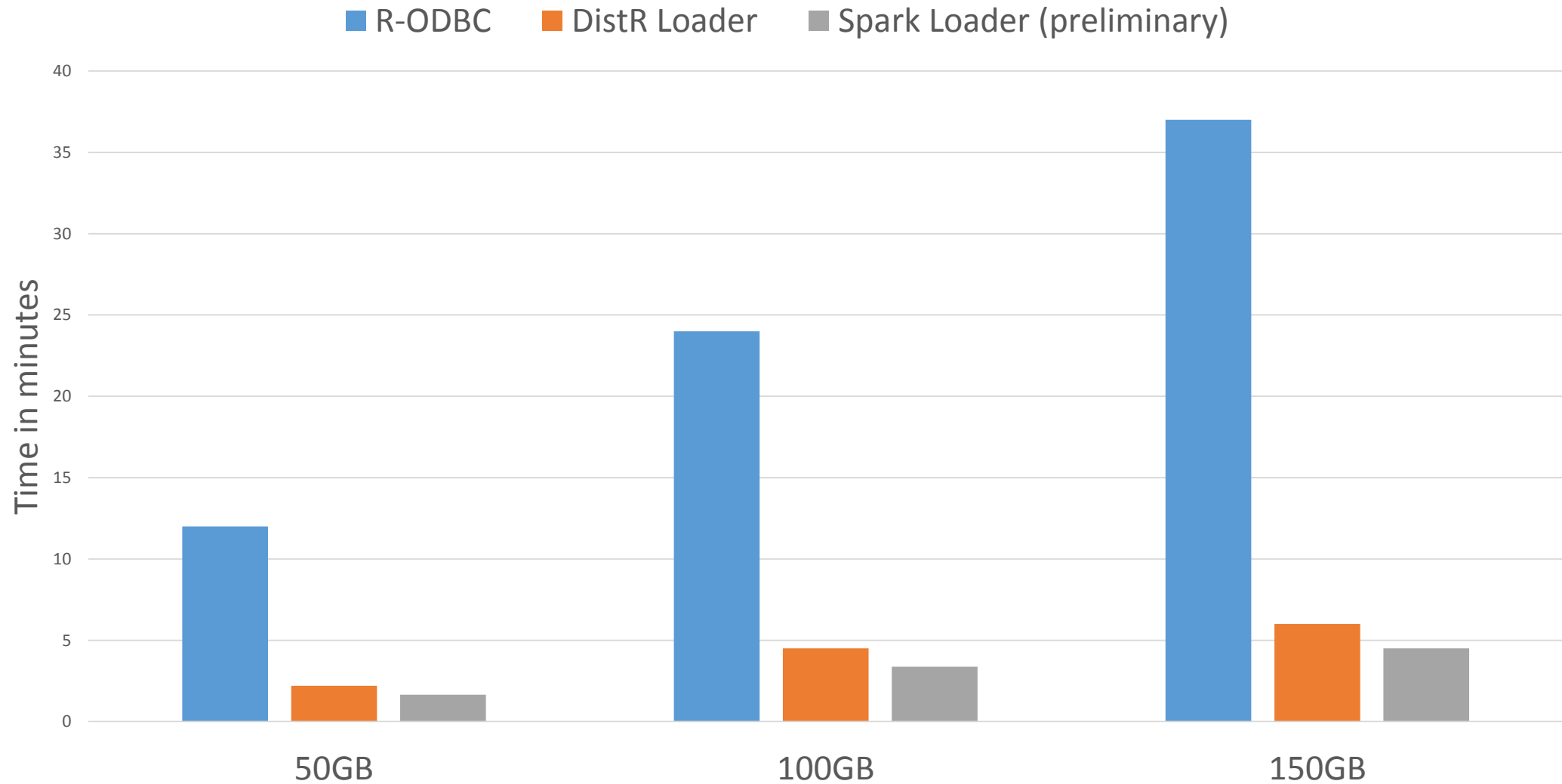
Uniform Distribution Policy



# Data loading from Vertica to DistR and Spark



# Data loading from Vertica to DistR and Spark



# Integration Status

- Distributed R
  - Open sourced: <https://github.com/vertica/DistributedR>
  - *Eurosys* 2013, *SIGMOD* 2015
  - K-Means, PageRank, Linear/Logistic Regression, Random Forest
- Spark, Caffe – Works In Progress

# Thank You!

## References

- Open Source Distributed R: <https://github.com/vertica/DistributedR>
- S. Venkataraman et al, *Presto: Distributed Machine Learning and Graph Processing with Sparse Matrices*, Eurosys 13.
- S. Prasad et al, *Large-scale Predictive Analytics in Vertica: Fast Data Transfer, Distributed Model Creation, and In-database Prediction*, SIGMOD 15 (to appear).

## Acknowledgements

- Distributed R team members: Shreya Prasad, Arash Fard, Vishrut Gupta, Jorge Martinez, Edward Ma, Indrajit Roy, Sunil Venkayala, Vincent Xu