Integrating Relational Data with Hadoop and Spark

think 2018

—

Stefan Hummel
stefan.hummel@de.ibm.com

Andreas Weininger
andreas.weininger@de.ibm.com

IBM

Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

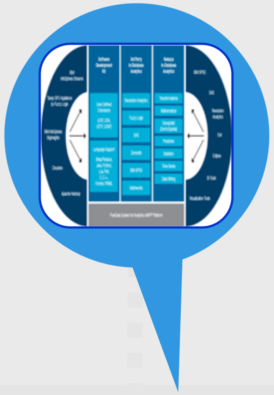
Labs

- Lab 1: Creating Tables
- Lab 2: Loading Data
- Lab 3: Executing Queries
- Lab 4: Linear Regression in SQL
- Lab 5: Linear Regression with R
- Lab 6: Linear Regression with Python
- Lab 7: Analyzing data with R in RStudio

Db2 Warehouse – Analytics Warehouse as a Service

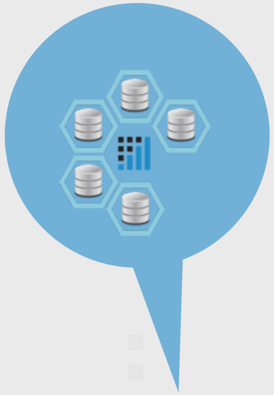
BLU Acceleration

- DB2 BLU columnar technology
- in-memory processing, data skipping, actionable compression, parallel vector processing, , “Load & Go” administration



Netezza In-Database Analytics

- Netezza predictive analytic algorithms
- fully integrated RStudio & R language



Db2 Warehouse MPP

- Massively Parallel Processing (MPP)
- Oracle compatibility
- fully-managed warehouse

Db2 Warehouse – Key Use Cases

Cloudant Analytics

- Easy synchronization of JSON to structured data
- Allows analytics via standard BI tools
- In-database predictive algorithms allow greater insight for Cloudant users than ever before

Extend or Modernize

- Extend on-premise data warehouses to the cloud
- Flexible, cost-effective growth
- Hybrid Cloud model to support ground to cloud

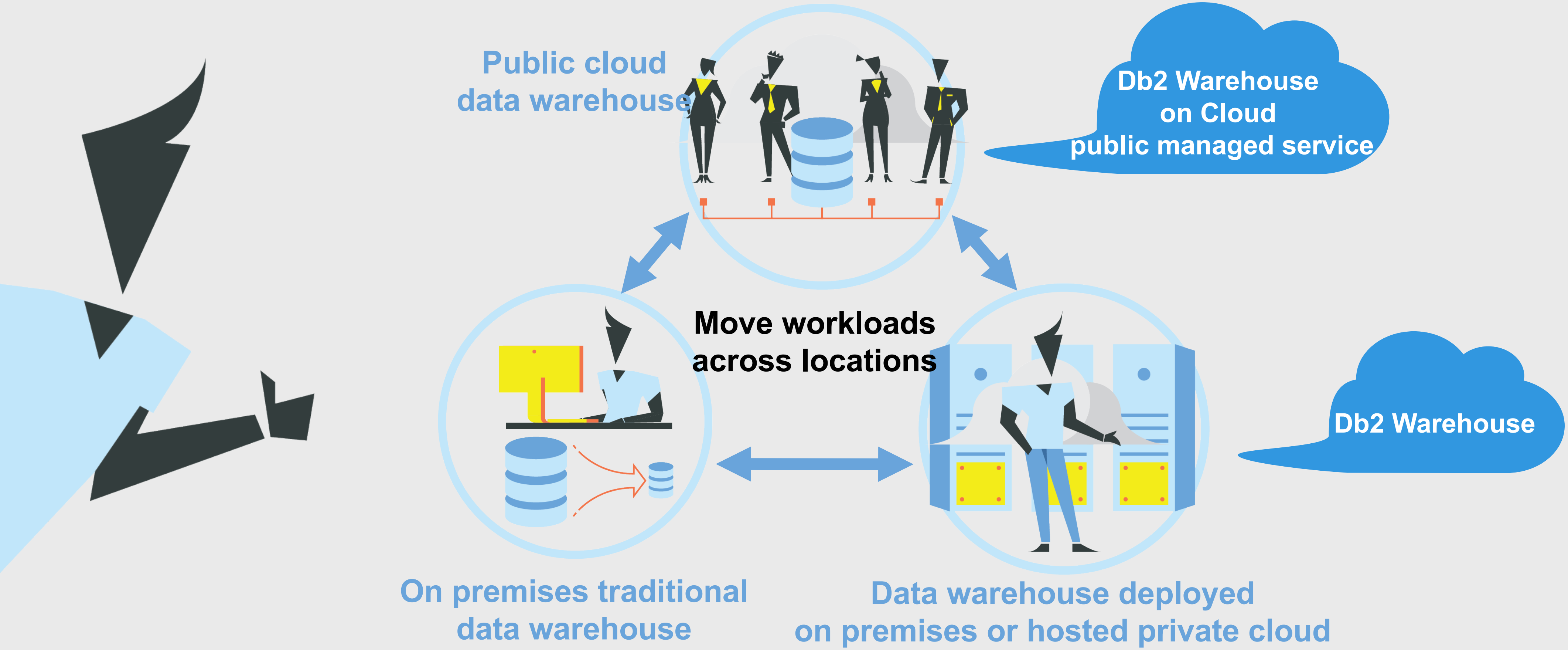
In-Database Analytics

- Robust predictive analytic algorithms
- Integrated with R (and Python/Spark)
- Watson Analytics Ready
- Analytics Ecosystem with Partners

Data Warehouse & Analytics Service

- Data Warehousing and Analytics in the Cloud
- Cloud Agility and Flexibility
- Analytics for Cloud Data, Data Marts, and development and test environments

Db2 Warehouse offerings – on cloud or local



Public cloud data warehouse

Db2 Warehouse on Cloud public managed service

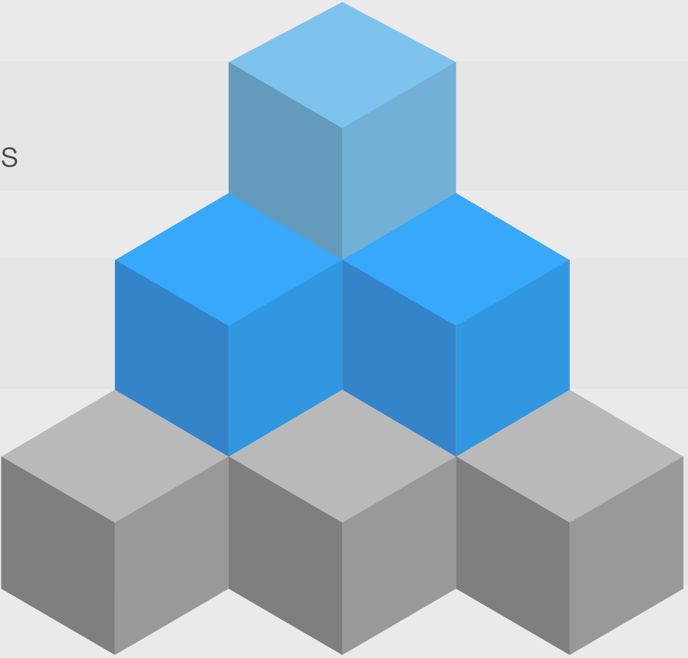
Move workloads across locations

Db2 Warehouse

On premises traditional data warehouse

Data warehouse deployed on premises or hosted private cloud

Db2 Warehouse – machine learning algorithms & geospatial



Decision Trees
Regression Trees

Naive Bayes

Linear Regression

k-Means Clustering
k-Nearest Neighbors

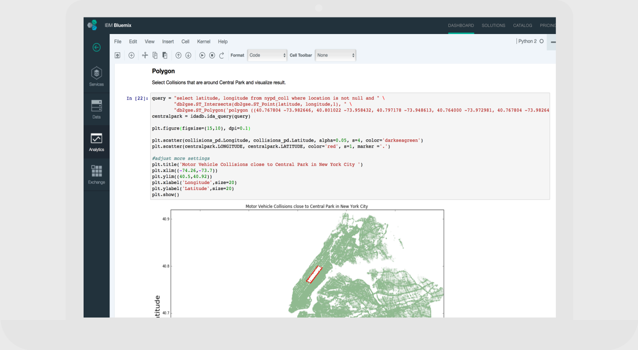
Assosiation Rules
Sequential Patters & Rules

Geospatial Extender

In-database analytics on Db2 Warehouse

Run machine learning algorithms and do geospatial analytics directly on the data in Db2 Warehouse leveraging the efficient parallel database engine.

Db2 Warehouse – R and Python integration



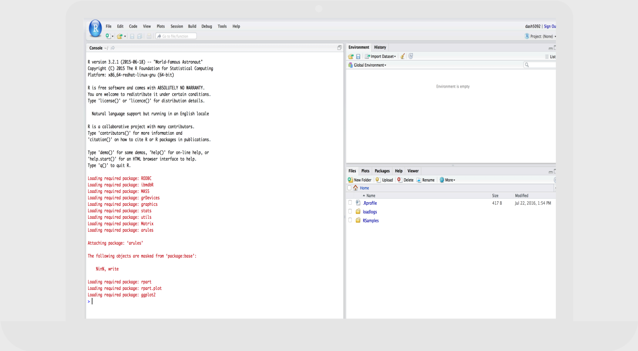
Polygon

Select Collisions that are around Central Park and visualize result.

```
In [22]: query = "select latitude, longitude from nypd_coll where location is not null and " \
                 "db2gse.ST_Intersects(db2gse.ST_Point(latitude, longitude,1), " \
                 "db2gse.ST_Polygon('polygon ((40.767804 -73.982646, 40.801022 -73.958432, 40.797178 -73.948613, 40.764000 -73.972981, 40.767804 -73.98264
         centralpark = idadb.ida_query(query)

         plt.figure(figsize=(15,10), dpi=0.1)

         plt.scatter(collisions_pd.Longitude, collisions_pd.Latitude, alpha=0.05, s=4, color='darkseagreen')
         plt.scatter(centralpark.LONGITUDE, centralpark.LATITUDE, color='red', s=1, marker ='.')

         #adjust more settings
         plt.title('Motor Vehicle Collisions close to Central Park in New York City ')
         plt.xlim((-74.26,-73.7))
         plt.ylim((40.5,40.92))
         plt.xlabel('Longitude',size=20)
         plt.ylabel('Latitude',size=20)
         plt.show()
```

Motor Vehicle Collisions close to Central Park in New York City
40.9
40.8
40.7
Latitude

- Integrated with Spark Service offering in Bluemix
- Access data in Db2 Warehouse from notebook interface in Spark Service or Data Science Experience GUI
- Run machine learning algorithms directly in database engine



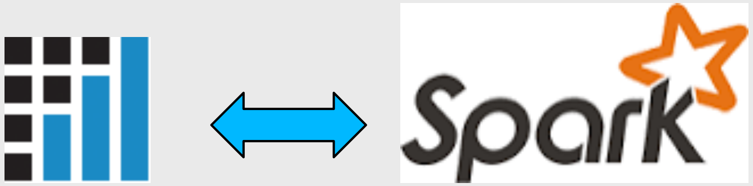
```
R version 3.2.1 (2015-06-18) -- "World-Famous Astronaut"
Copyright (C) 2015 The R Foundation for Statistical Computing
Platform: x86_64-redhat-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

Loading required package: RODBC
Loading required package: ibmdbR
Loading required package: MASS
Loading required package: grDevices
Loading required package: graphics
Loading required package: stats
Loading required package: utils
Loading required package: Matrix
Loading required package: arules

Attaching package: 'arules'

The following objects are masked from 'package:base':

    %in%, write

Loading required package: rpart
Loading required package: rpart.plot
Loading required package: ggplot2
>
```

Name	Size	Modified
.Rprofile	417 B	Jul 22, 2016, 1:54 PM
loadlogs		
RSamples		

- Access data in Db2 Warehouse from RStudio environment
- RStudio hosted on Db2 Warehouse server
- Machine learning algorithms of Db2 Warehouse available via ibmdbR library
- Possibility to pushdown R code to Db2 Warehouse server

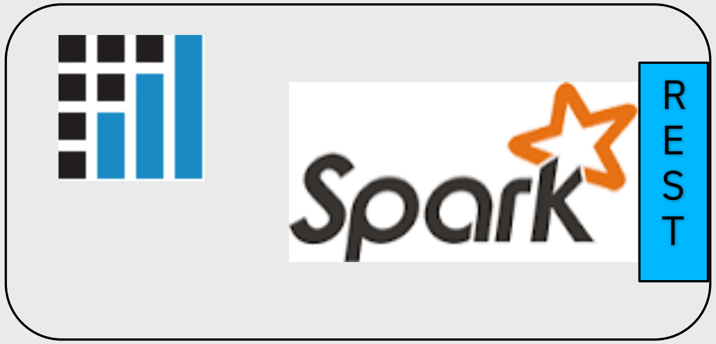
Db2 Warehouse – Spark Integration

Communication



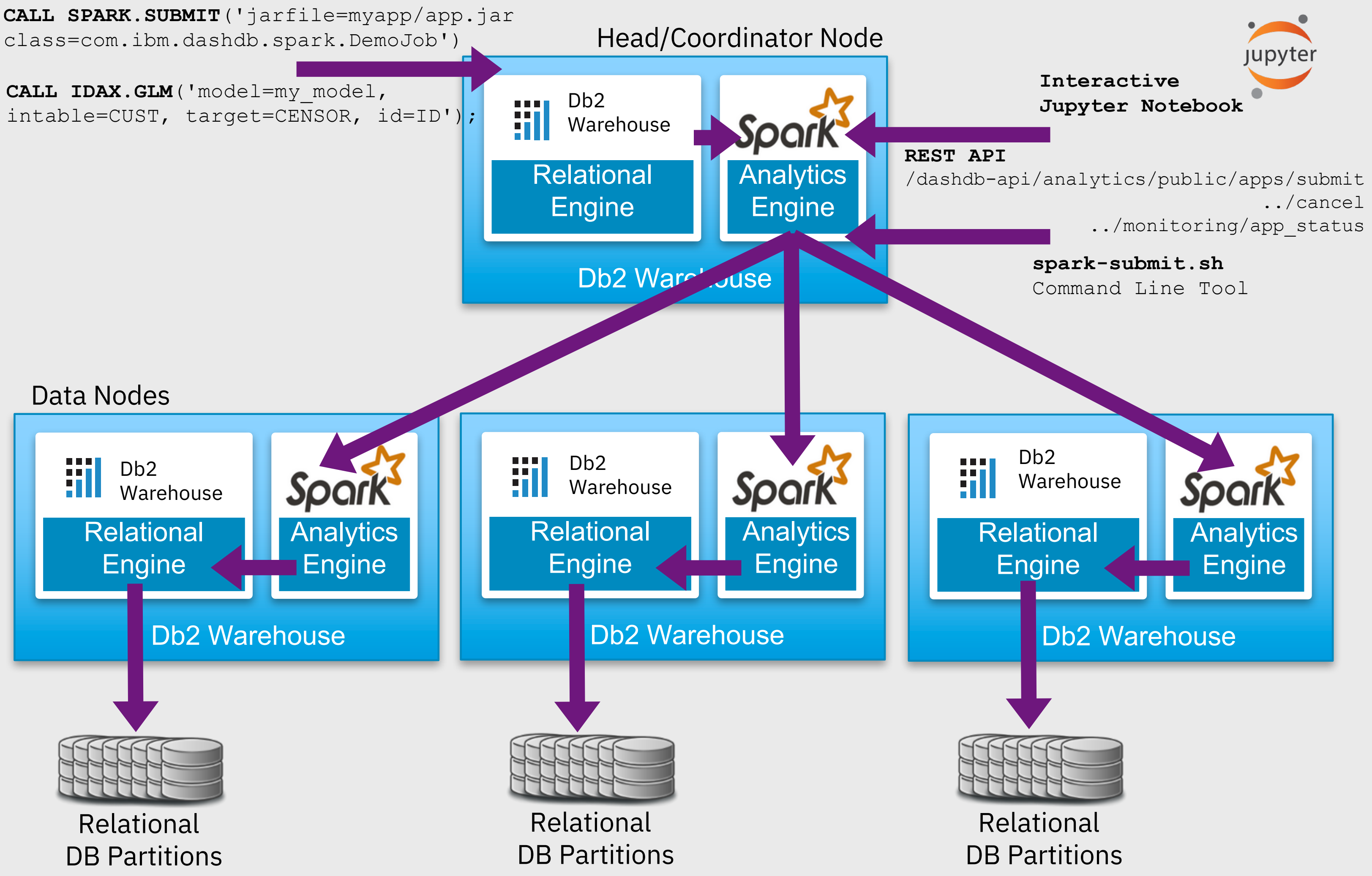
- Spark reading and writing Db2 Warehouse tables

Administration



- Administration of Spark code via
  - •REST interface
- Monitoring of Spark via
  - •Db2 Warehouse Web Console or
  - •REST Interface

Db2 Warehouse Spark Integration Examples



```
CALL SPARK.SUBMIT('jarfile=myapp/app.jar
class=com.ibm.dashdb.spark.DemoJob')

CALL IDAX.GLM('model=my_model,
intable=CUST, target=CENSOR, id=ID');
```

Interactive Jupyter Notebook

REST API

```
/dashdb-api/analytics/public/apps/submit
                                ../cancel
                    ../monitoring/app_status
```

spark-submit.sh
Command Line Tool

Some hints for executing the labs

- Start with starting everything as described in in Appendix I
- Check with

  ```
  docker exec Db2wh status
  ```

  whether Db2 Warehouse is running
- Remove the directory Advanced-Analytics in the home directory of user ibmuser with

  ```
  rm –rf Advanced-Analytics
  ```
- Only after that execute the `git clone` command
- The Jupyter notebooks can be found in the subdirectory Advanced-Analytics/notebooks

Notices and disclaimers

U.S. Government Users Restricted Rights — use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM.

Information in these presentations (including information relating to products that have not yet been announced by IBM) has been reviewed for accuracy as of the date of initial publication and could include unintentional technical or typographical errors. IBM shall have no responsibility to update this information. **This document is distributed "as is" without any warranty, either express or implied. In no event, shall IBM be liable for any damage arising from the use of this information, including but not limited to, loss of data, business interruption, loss of profit or loss of opportunity.** IBM products and services are warranted per the terms and conditions of the agreements under which they are provided.

IBM products are manufactured from new parts or new and used parts. In some cases, a product may not be new and may have been previously installed. Regardless, our warranty terms apply."

Any statements regarding IBM's future direction, intent or product plans are subject to change or withdrawal without notice.

Performance data contained herein was generally obtained in a controlled, isolated environments. Customer examples are presented as illustrations of how those
customers have used IBM products and the results they may have achieved. Actual performance, cost, savings or other results in other operating environments may vary.

References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business.

Workshops, sessions and associated materials may have been prepared by independent session speakers, and do not necessarily reflect the views of IBM. All materials and discussions are provided for informational purposes only, and are neither intended to, nor shall constitute legal or other guidance or advice to any individual participant or their specific situation.

It is the customer's responsibility to insure its own compliance with legal requirements and to obtain advice of competent legal counsel as to the identification and interpretation of any relevant laws and regulatory requirements that may affect the customer's business and any actions the customer may need to take to comply with such laws. IBM does not provide legal advice or represent or warrant that its services or products will ensure that the customer follows any law.

Notices and disclaimers continued

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products about this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products. IBM does not warrant the quality of any third-party products, or the ability of any such third-party products to interoperate with IBM's products. **IBM expressly disclaims all warranties, expressed or implied, including but not limited to, the implied warranties of merchantability and fitness for a purpose.**
The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents, copyrights, trademarks or other intellectual property right.

IBM, the IBM logo, ibm.com and [names of other referenced IBM products and services used in the presentation] are trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at: www.ibm.com/legal/copytrade.shtml.
.

Thank you

Stefan Hummel	Andreas Weininger
—	
stefan.hummel@de.ibm.com	andreas.weininger@de.ibm.com
+49-160-742-1795	+49-172-756-5266
ibm.com	