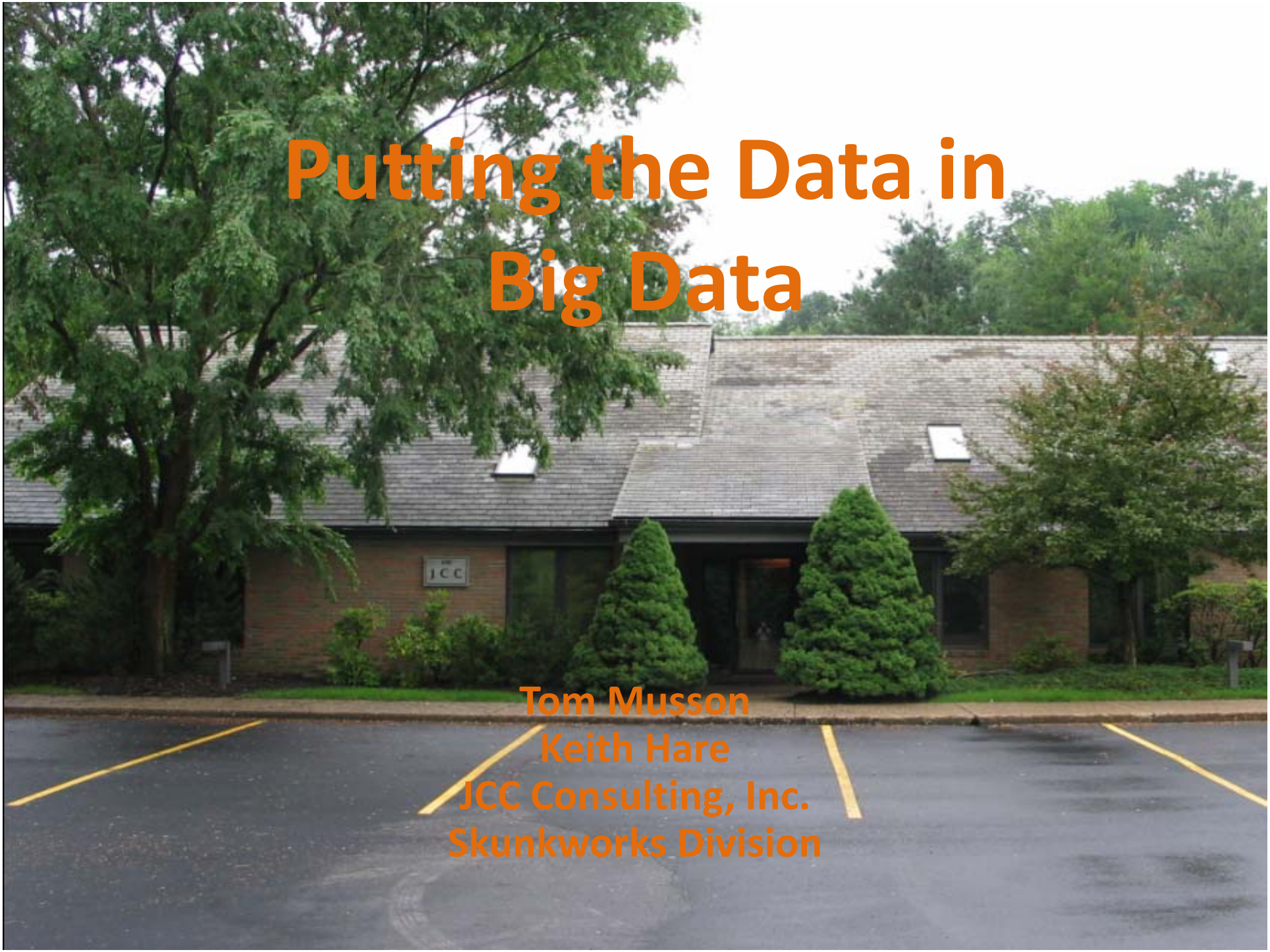


Putting the Data in Big Data

Tom Musson
Keith Hare
JCC Consulting, Inc.
Skunkworks Division

Introduction

- JCC LogMiner Loader is a flexible tool for replicating Rdb changes to other databases
- Can replicate to a number of targets
- Look at two targets
  - MongoDB
  - HP Vertica
- Skunkworks testing

Sources and Targets

- Source
  - Oracle Rdb (any version that supports the LogMiner)
- Target
  - Oracle Rdb (any version that supports multi-statement procedures)
  - Oracle (requires SQL*net client on the system running the Loader)
    - 12.1
    - 11.0, 11.1
    - 10.1, 10.2
    - 9.0, 9.2
    - 8.1 - deprecated
    - Oracle client-server compatibility (see Oracle matrix)
  - JDBC Class 4 driver
  - Tuxedo
  - XML (to your own API)
  - File

Sources and Targets (cont.)

- The source and target can be different.
  - Format of the target can be completely different than the format of the source database.
    - Logically
    - Physically
  - Tuning of the target can be different.
    - Indices
    - Placement
    - Buffering
    - Caching
- But how different can the target be? What about NoSQL (or Not-only) SQL databases?

NoSQL Databases

- NoSQL Marketing Position
  - Minimal upfront design
  - Few common features
  - Horizontally Scalable – Big Data
- Look at two
  - MongoDB – question at the 2013 Rdb Forum in Paris
  - HP Vertica – JDBC driver is available

What do I do with Big Data?

- Analytics
  - Capacity planning
  - Quality of Service
  - Customer Personalization
  - Predict chronic illness
- Web Interface
- Data Integration

- Tools
  - JSON Studio
  - Pentaho
  - Etc.

MongoDB

- An open-source document database
  - Open Source version: www.mongodb.org
  - Production support: www.mongodb.com
  - We are running MongoDB on 64-bit Linux
- Stores and retrieves JSON documents in collections (“tables”)
- Schema-less – no upfront design required

JDBC Driver for MongoDB

- Downloaded from UnityJDBC
  - http://www.unityjdbc.com/mongojdbc/mongo_jdbc.php
- Layers on MongoDB Java Driver
- Variation 1 – separate JAR files for Java driver and JDBC layer
  - mongo-java-driver-2-12-2.jar
  - mongodb_unityjdbc.jar
- Variation 2 – both combined into a single JAR file
  - mongodb_unityjdbc_full.jar

JSON – Java Script Object Notation

- Employees Example:

```
{"EMPLOYEE_ID"    :"00160"
,"LAST_NAME"      :"Hare"
,"FIRST_NAME"     :"Keith"
,"MIDDLE_INITIAL" :"W"
,"ADDRESS_DATA_1" :"600 Newark Road"
,"ADDRESS_DATA_2" :"P.O. Box 381"
,"CITY"           :"Granville"
,"STATE"          :"OH"
,"POSTAL_CODE"    :"43023"
,"SEX"            :"M"
,"BIRTHDAY"       : new Date()
,"STATUS_CODE"    :"1"
}
```

- But where are the datatypes?
- Similar to XML but without the rigor

MongoDB Command Line

- Logging into a MongoDB database

```
[keith@jcc-rh-oms ~]$ mongo
MongoDB shell version: 2.6.4
connecting to: test
> db.auth("keith", "mypassword")
1
```

- Linux terminal session
- Result of function (method) db.auth is 1

Insert a Document

Assign Document to Variable

```
> post =
{"EMPLOYEE_ID"    :"00161"
,"LAST_NAME"      :"Musson"
,"FIRST_NAME"     :"Tom"
,"MIDDLE_INITIAL" :"H"
,"ADDRESS_DATA_1" :"600 Newark Road"
,"ADDRESS_DATA_2" :"P.O. Box 381"
,"CITY"           :"Granville"
,"STATE"          :"OH"
,"POSTAL_CODE"    :"43023"
,"SEX"            :"M"
,"BIRTHDAY"       : new Date()
,"STATUS_CODE"    :"1"
}
>
```

Contents of variable are echoed

```
{
    "EMPLOYEE_ID" : "00161",
    "LAST_NAME" : "Musson",
    "FIRST_NAME" : "Tom",
    "MIDDLE_INITIAL" : "H",
    "ADDRESS_DATA_1" : "600 Newark Road",
    "ADDRESS_DATA_2" : "P.O. Box 381",
    "CITY" : "Granville",
    "STATE" : "OH",
    "POSTAL_CODE" : "43023",
    "SEX" : "M",
    "BIRTHDAY" : ISODate("2014-08-
19T13:58:18.861Z"),
    "STATUS_CODE" : "1"
}
```

Insert the document into collection

```
> db.employees.insert(post)
WriteResult({ "nInserted" : 1 })
>
```

Retrieving a Document

- Find all of the documents in a collection

```
> db.employees.find()
{ "_id" : ObjectId("53f357fa173a5903c1f37744"), "EMPLOYEE_ID" : "00161",
"LAST_NAME" : "Musson", "FIRST_NAME" : "Tom", "MIDDLE_INITIAL" : "H",
"ADDRESS_DATA_1" : "600 Newark Road", "ADDRESS_DATA_2" : "P.O. Box 381",
"CITY" : "Granville", "STATE" : "OH", "POSTAL_CODE" : "43023", "SEX" :
"M", "BIRTHDAY" : ISODate("2014-08-19T13:58:18.861Z"), "STATUS_CODE" :
"1" }
>
```

- "_id" : ObjectId("53f357fa173a5903c1f37744")
  - MongoDB adds a unique ObjectID
  - Builds an index on the ObjectIDs in each document collection

Retrieving a Document

- Find a specific employees record:

```
> db.employees.find({"EMPLOYEE_ID" : "00173"})
{ "_id" : ObjectId("53f3722d49c6ffeadef760db"), "LAST_NAME" : "Bartlett
", "FIRST_NAME" : "Dean      ", "MIDDLE_INITIAL" : "G", ADDRESS_DATA_1" :
"149 Steeple Lane         ", "ADDRESS_DATA_2" : "                    ",
"CITY" : "Troy               ", "STATE" : "NH", "POSTAL_CODE" : "03465",
"SEX" : "M", "BIRTHDAY" : ISODate("1927-03-05T00:00:00Z"), "STATUS_CODE"
: "1", "EMPLOYEE_ID" : "00173" }
>
```

- Predicate is a JSON document
- “Column names” are case sensitive
- JCC LogMiner Loader could be configured to trim trailing spaces

But where are the datatypes?

- Loosely typed
  - Character strings in quotes
  - Numerics without quotes
  - ISODate()
  - ObjectId (_id ‘column’)
- Documents in a collection can have different structures

FTP JDBC Driver to OpenVMS

- OpenVMS

```
Directory JCC_ROOT:[KEITH.MONGODB.2014-08-17]
mongo-java-driver-2-12-2.jar;1
                1152/1155     17-AUG-2014 23:26:49.98  (RWED,RWED,RE,)
mongodb_unityjdbc.jar;2
                2564/2565     18-AUG-2014 08:16:13.13  (RWED,RWED,RE,)
mongodb_unityjdbc_full.jar;3
                3714/3715     18-AUG-2014 08:16:16.17  (RWED,RWED,RE,)
Total of 3 files, 7430/7435 blocks.
```

- Reset the Jar File characteristics

```
pandor > set file/attr=(rfm:stmlf,rat:cr,lrl:0,mr:0) *.jar;*/log
```

Loader Configuration File

```
!
! MongoDB
!
output~jdbc~synch~jdbc:mongo://jcc-rh-oms.jcc.com:27017/test?rebuildschema=true
jdbc~connect~jdbc:mongo://jcc-rh-oms.jcc.com:27017/test?rebuildschema=true
!
validation~keith~mypassword
!
!    Affinity MongoDB JDBC Driver
!
jdbc~classpath~/jcc_root/keith/mongodb/2014-08-17/mongo-java-driver-2-12-2.jar
jdbc~classpath~/jcc_root/keith/mongodb/2014-08-17/mongodb_unityjdbc.jar

!
!  …or the unified driver
!
jdbc~classpath~/jcc_root/keith/mongodb/2014-08-17/mongodb_unityjdbc_full.jar

!
! The remainder of this file is standard Loader "stuff"
```

MongoDB Connect String

```
jdbc:mongo://jcc-rh-oms.jcc.com:27017/test?rebuildschema=true
```

- MongoDB installed on node jcc-rh-oms.jcc.com
- MongoDB Port: 27017
- /test?rebuildschema=true
  - Schema on demand – rebuilds a relational schema at database attach time
  - Collections (“tables”) created after attach not visible until next rebuildschema

MongoDB Database Prep

- JCC LogMiner Loader maps source tables to target tables
- The Unity MongoJDBC Driver maps tables to MongoDB collections
- Target tables (collections) have to exist at loader startup
- Create by inserting a single JSON document for each table with appropriate columns

Target Characteristics

- JCC LogMiner Loader supports transactional consistency
- MongoDB supports statement level atomicity
- A failure during the application of a multi-statement source transaction might have applied only part of a source transaction to the target
  - Restarting the Loader will write the complete transaction

MongoDB Performance

- Replicate option does a lot of reading
- Possible to define indexes:

```
> db.employees.ensureIndex( { "EMPLOYEE_ID": 1})
{ "ok" : 1 }
>
> db.salary_history.ensureIndex( {"EMPLOYEE_ID":1, "SALARY_START":1} )
{ "ok" : 1 }
>
```

- Improves performance of updates
- Unsure about impact on inserts

Index on EmployeeID helps

```
Rate:  10.00                        KWH_MONGO              19-SEP-2014 11:20:32.21
================================================================================
   Input: 19-SEP-2014 11:20:21.66                Output: 19-SEP-2014 11:20:21.63
--[Trail:   10.56]---------------              ---[Trail:   10.58]---------------
Transactions                   66              Checkpoints                     56
Records                    156767                 Timeout                      0
   Modify                  146601                 BufferLimit(  3958)          1
   Delete                   10100                 NoWork                       0
   Commit                      66              Records(   1)               147855
 Discarded                                        Messages(  N/A )           N/A
   Filtered                     0                 Filtered                     0
   Excluded                     0              Failure                         0
   Unknown                      0              Timeout                         0
   Restart                   8801         - Current ---------------- Ave/Second -
   NoWork                       8         Checkpoints            37         3.70
   Heartbeat                    0         Records              3656       365.60
Timeout                        58         Rate                 4.97%
--- Restart Context ------                - Latency(sec) ------ LML detail ------
M|AIJ#                       20 |          CLM    5.52 |  Inpt   1.8%  Cnvt  15.9%
Q|VBN                     2234 |          ------------   Sort   0.0%  Trgt  82.1%
P|TSN                     6429 |          LML    0.27    Sync   0.0%  Ckpt   0.1%
 CTSN                     6429 |       - Loaders - 0 -------------------------------
  LSN                      303 |       - States  - >
```

HP Vertica Community Edition

- Downloaded a development virtual appliance
  - Prebuilt Linux system with Vertica installed
  - Web site https://my.vertica.com/community/
  - Import as a VMware virtual appliance
- Vertica WEB UI

  https://vertica.jcc.com:5450/webui/
- Admin Tools – logged into Linux as DBAdmin

```
[dbadmin@vertica keith]$ admintools
```

What is Vertica?

- Column store database
  - Columns for a table are stored together rather than rows
  - Performance benefit when part of the columns are accessed
- Horizontally scalable
- High availability
- SQL interface
- Analysis tools

HP Vertica Drivers

- My.vertica.com downloads – login required
- HP Vertica 7.0.2 Client Packages for the Community Edition
  - JDBC Driver
    - Installs to

      C:\Program Files\Vertica Systems\JDBC\ vertica-jdbc-7.0.2-1.jar
    - I created a copy with underscores instead of periods
  - ODBC
  - ADO.net

FTP JDBC Driver to OpenVMS

- OpenVMS

```
Directory JCC_ROOT:[KEITH.VERTICA]
vertica-jdbc-7_0_2-1.jar;1
                1436/1440     20-AUG-2014 16:21:43.74   (RWED,RWED,RE,)

Total of 1 file, 1436/1440 blocks.
```

- Reset the Jar File characteristics

```
pandor > set file/attr=(rfm:stmlf,rat:cr,lrl:0,mr:0) *.jar;*/log
```

Loader Configuration File

```
!
! Vertica
!
output~jdbc~synch~jdbc:vertica://vertica.jcc.com:5433/JCCVertica
jdbc~connect~jdbc:vertica://vertica.jcc.com:5433/JCCVertica
!
validation~keith~<password>
!
!    HP Vertica JDBC Driver
!
jdbc~driver~com.vertica.jdbc.Driver
jdbc~classpath~/jcc_root/keith/vertica/vertica-jdbc-7_0_2-1.jar
!
! The remainder of this file is standard Loader "stuff"
!
```

Vertica Connect String

```
jdbc:vertica://vertica.jcc.com:5433/JCCVertica
```

- Vertica installed on node vertica.jcc.com
- Vertica Port: 5433
- Database name: JCCVertica

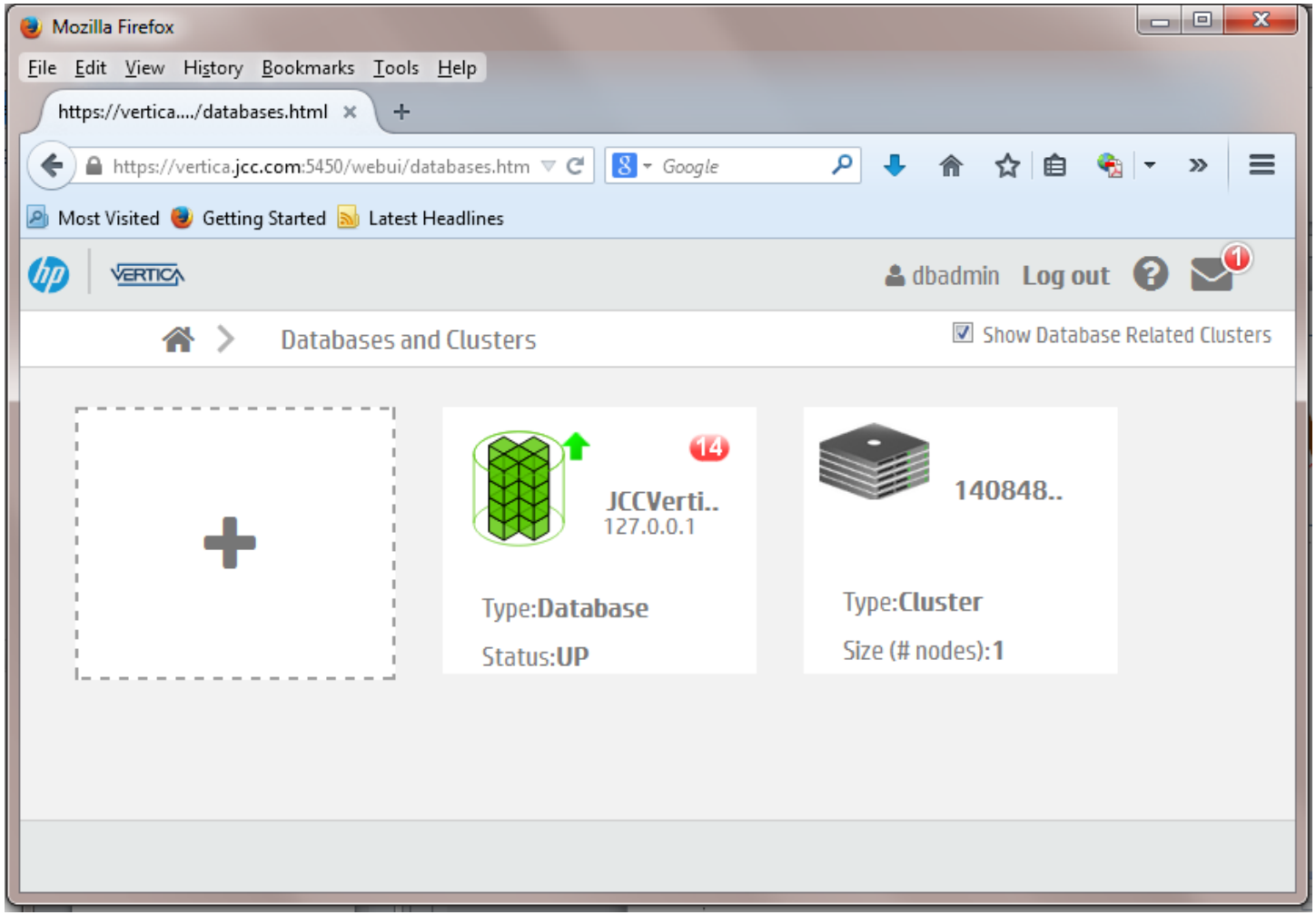
Target Characteristics

- JCC LogMiner Loader supports transactional consistency
- Vertica supports multi-statement transactions
- A failure during the application of a multi-statement source transaction will be rolled back
  - The Loader will restart where it left off

Vertica Database Prep

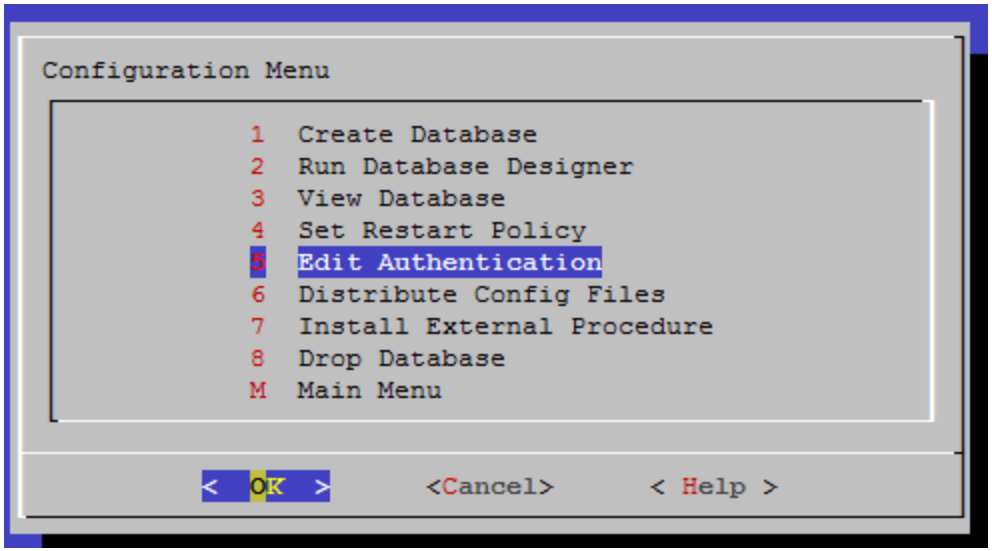
- Create Database
- Allow Access from subnet
- Create tables with a script
- View the schema and tables

Create Database with Web UI



Vertica Admintools to Allow Access

- Admintools Configuration Menu



- To access from other nodes, add entry such as:

```
ClientAuthentication = host all 172.16.0.0/16 password
```

Use VSQL to Create Tables

VSQL \i command executes a script:

```
[dbadmin@vertica ~]$ vsql
Password:
Welcome to vsql, the Vertica Analytic Database interactive terminal.

Type:  \h or \? for help with vsql commands
       \g or terminate with semicolon to execute query
       \q to quit

dbadmin-> \i Vertica_Personnel_DB_Script_2014-08-20.sql
CREATE SCHEMA
GRANT PRIVILEGE
CREATE TABLE
ALTER TABLE
…
```

VSQL Commands - Schemas

VSQL \dn command shows all schemas:

```
dbadmin=> \dn
        List of schemas
    Name     |   Owner   | Comment
-------------+-----------+---------
 v_internal  | dbadmin   |
 v_catalog   | dbadmin   |
 v_monitor   | dbadmin   |
 public      | dbadmin   |
 personnel   | dbadmin   |
(5 rows)
```

VSQL Commands - Tables

VSQL \dt command lists accessible tables:

```
dbadmin=> \dt
                  List of tables
 Schema |       Name       | Kind  |  Owner   | Comment
--------+------------------+-------+----------+---------
 public | CANDIDATES       | table | dbadmin  |
 public | COLLEGES         | table | dbadmin  |
 public | DEGREES          | table | dbadmin  |
 public | DEPARTMENTS      | table | dbadmin  |
 public | EMPLOYEES        | table | dbadmin  |
 public | JOBS             | table | dbadmin  |
 public | JOB_HISTORY      | table | dbadmin  |
 public | RESUMES          | table | dbadmin  |
 public | SALARY_HISTORY   | table | dbadmin  |
 public | WORK_STATUS      | table | dbadmin  |
(10 rows)
```

VSQL Commands - Describe

VSQL \d command describes a table:

```
dbadmin=> \d employees
                                List of Fields by Tables
 Schema |   Table   |     Column      |   Type   | Size | Default | Not Null | Primary Key | Foreign Key
--------+-----------+-----------------+----------+------+---------+----------+-------------+-------------
 public | EMPLOYEES | EMPLOYEE_ID     | char(5)  |    5 |         | t        | t           |
 public | EMPLOYEES | LAST_NAME       | char(14) |   14 |         | f        | f           |
 public | EMPLOYEES | FIRST_NAME      | char(10) |   10 |         | f        | f           |
 public | EMPLOYEES | MIDDLE_INITIAL  | char(1)  |    1 |         | f        | f           |
 public | EMPLOYEES | ADDRESS_DATA_1  | char(25) |   25 |         | f        | f           |
 public | EMPLOYEES | ADDRESS_DATA_2  | char(20) |   20 |         | f        | f           |
 public | EMPLOYEES | CITY            | char(20) |   20 |         | f        | f           |
 public | EMPLOYEES | STATE           | char(2)  |    2 |         | f        | f           |
 public | EMPLOYEES | POSTAL_CODE     | char(5)  |    5 |         | f        | f           |
 public | EMPLOYEES | SEX             | char(1)  |    1 |         | f        | f           |
 public | EMPLOYEES | BIRTHDAY        | date     |    8 |         | f        | f           |
 public | EMPLOYEES | STATUS_CODE     | char(1)  |    1 |         | f        | f           |
(12 rows)
```

Vertica Performance Tuning

- Look at:

```
JCCVertica-> ANALYZE_STATISTICS('JCCVertica')
JCCVertica-> ANALYZE_HISTOGRAM('JCCVertica')
JCCVertica-> ANALYZE_WORKLOAD('')
```

- Vertica Management Console
  - Database Designer
- We have more to learn about Vertica Tuning

Vertica Tuning needs more effort

```
Rate:  10.00                        KWH_VERTICA            21-SEP-2014 17:30:28.05
================================================================================
   Input: 21-SEP-2014 17:26:04.54                Output: 21-SEP-2014 17:26:04.50
--[Trail:    4.4m]---------------              ---[Trail:    4.4m]---------------
Transactions                   72              Checkpoints                    69
Records                     14313                 Timeout                      1
   Modify                    7000                 BufferLimit(   150)          1
   Delete                    7241                 NoWork                       0
   Commit                      72              Records(   1)                7000
 Discarded                                        Messages(  N/A )            N/A
   Filtered                     0                 Filtered                     0
   Excluded                     0              Failure                         0
   Unknown                      0              Timeout                         0
   Restart                   7242          - Current ---------------- Ave/Second -
   NoWork                       1          Checkpoints              3        0.30
   Heartbeat                    0          Records                300       30.00
Timeout                         7          Rate                 1.33%
--- Restart Context ------                 - Latency(sec) ------ LML detail ------
M|AIJ#                     22  |            CLM    4.2m  |  Inpt   0.1%  Cnvt   0.7%
Q|VBN                   6154  |            ------------   Sort   0.0%  Trgt  99.2%
P|TSN                   7402  |            LML    3.62    Sync   0.0%  Ckpt   0.0%
 CTSN                   7402  |         - Loaders - 0 ------------------------------
  LSN                    483  |         - States  - z
```

Tuning to HP Vertica Strengths

Vertica is a column-store so is likely to perform better if it gets multiple records at a time

- $ define jcc_lml_jdbc_batch_size 50
- Increase checkpoint interval

Examples use Targets Sub-optimally

- Both examples use Replicate

```
MapTable~EMPLOYEES~EMPLOYEES,employees~Replicate
MapColumn~EMPLOYEES~EMPLOYEE_ID
...
```

- Vertica & MongoDB optimized for inserts
- Consider AUDIT instead of Replicate
  - Updates and Deletes inserted in Target
  - Add Virtual columns
    - TSN, TRANSACTION_COMMIT_TIME, TRANSACTION_START_TIME, ACTION, JCCLML_USERNAME

MongoDB Using Audit

```
Rate:  10.00                          KWH_MONGO              19-SEP-2014 12:14:31.80
================================================================================
   Input: 19-SEP-2014 12:14:20.54                Output: 19-SEP-2014 12:14:20.50
--[Trail:   11.26]---------------             ---[Trail:   11.30]---------------
Transactions                    90            Checkpoints                      86
Records                      39090               Timeout                        1
   Modify                    28900               BufferLimit(   499)            1
   Delete                    10100               NoWork                         0
   Commit                       90            Records(   1)                 28774
 Discarded                                        Messages(  N/A )             N/A
   Filtered                      0               Filtered                       0
   Excluded                      0            Failure                           0
   Unknown                       0            Timeout                           0
   Restart                   10201         - Current ---------------- Ave/Second -
   NoWork                        2         Checkpoints            64          6.40
   Heartbeat                     0         Records              6388        638.79
Timeout                         13         Rate                17.32%
--- Restart Context ------                 - Latency(sec) ------ LML detail ------
M|AIJ#                    21  |            CLM    7.08  |  Inpt   3.1%  Cnvt  25.8%
Q|VBN                  6805  |            ------------   Sort   0.0%  Trgt  70.8%
P|TSN                  6943  |            LML    0.16    Sync   0.1%  Ckpt   0.1%
 CTSN                  6943  |         - Loaders - 0 ------------------------------
  LSN                   553  |         - States  - >
```

Vertica Using Audit

```
Rate:  10.00                          KWH_VERTICA              21-SEP-2014 17:51:44.64
================================================================================
   Input: 21-SEP-2014 17:51:29.21                Output: 21-SEP-2014 17:51:29.15
--[Trail:   15.43]---------------              ---[Trail:   15.48]---------------
Transactions                    85              Checkpoints                     81
Records                      18385                 Timeout                       1
   Modify                     8200                 BufferLimit(   334)           1
   Delete                    10100                 NoWork                        0
   Commit                       85              Records(   1)                18150
 Discarded                                          Messages(  N/A )            N/A
   Filtered                      0                 Filtered                      0
   Excluded                      0              Failure                          0
   Unknown                       0              Timeout                          0
   Restart                     101           - Current ---------------- Ave/Second -
   NoWork                        2           Checkpoints            51         5.10
   Heartbeat                     0           Records              5050       505.00
Timeout                         13           Rate                16.42%
--- Restart Context ------                   - Latency(sec) ------ LML detail ------
M|AIJ#                    22  |              CLM   11.29 |  Inpt   2.5%  Cnvt  23.2%
Q|VBN                 15534  |              ------------   Sort   0.0%  Trgt  74.2%
P|TSN                  7520  |              LML    0.20    Sync   0.1%  Ckpt   0.0%
 CTSN                  7520  |           - Loaders - 0 -------------------------------
  LSN                   597  |           - States  - >
```

Other Possible Big Data Targets

- Oracle Big Data
- MarkLogic
- Others?

Oracle Big Data

- Cloudera Distribution of Hadoop (CDH)
  - Hive SQL Interface
    - Read only
    - Write to Hadoop with program interface
- Oracle NoSQL database
  - Based on BerkleyDB – a Java database
  - No JDBC interface
- Big Data Connectors
  - In Oracle, create a Linked Table interface to external document store
  - Read only

MarkLogic

- XML database
- Java interface
- Have not found a JDBC driver
- Custom interface possible, but would require code, test, and documentation

Future

- Operate using source names and datatypes
  - Support MongoDB without having to pre-create collections
- Support for other Big Data targets
- Find one of us to make suggestions or…
- Loader Suggestion Box
  - Jcc-lmloader@jcc.com

Summary

- JCC LogMiner Loader works with some Big Data targets
  - Need a JDBC Driver
  - Target performance tuning still needed
- Use the Big Data target that is tailored to your Big Data requirements

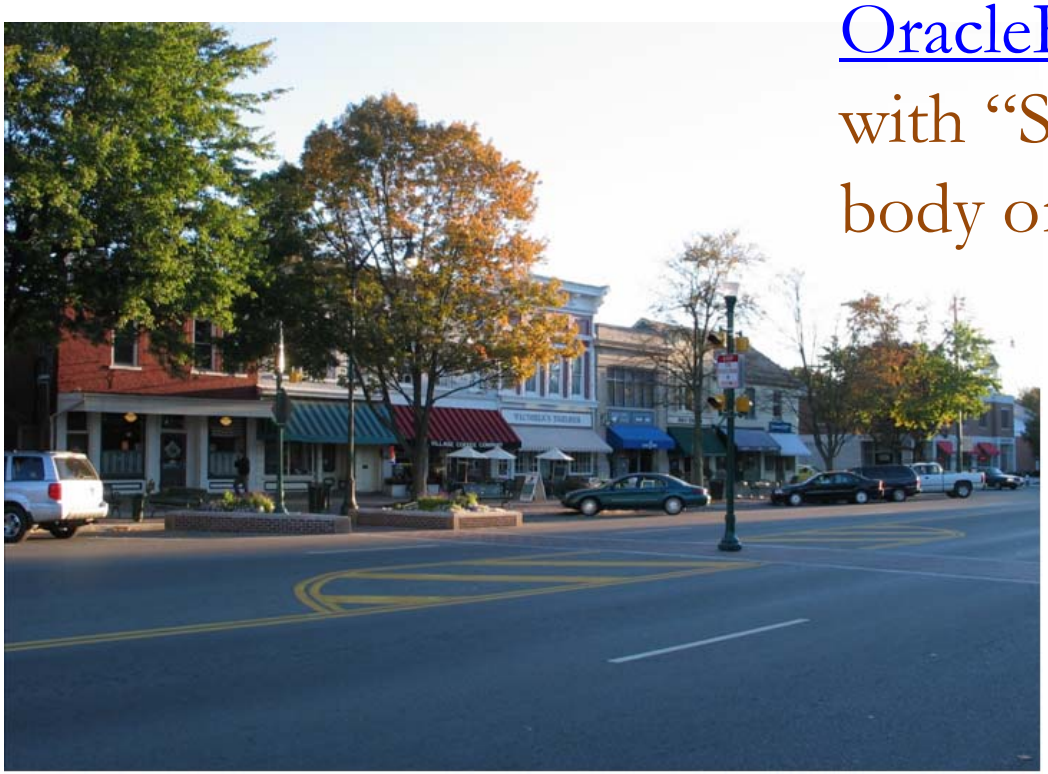
Acknowledgements

- Thanks to Rdb engineering for their support and counsel
- Thanks to our customers for sharing their experiences with the Loader

Join the Conversation

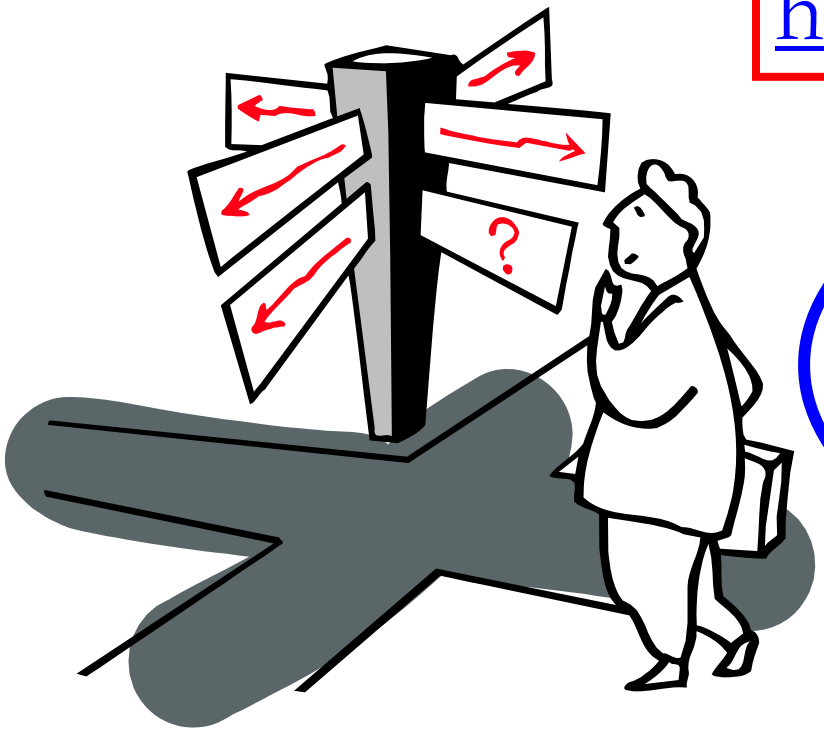
Join the worldwide Rdb community. Send mail to

OracleRdb-request@JCC.com with "SUBSCRIBE" in the body of the message.



Questions?

http://www.jcc.com/lml



At break,
please ask questions
and share ideas

Send your input and requests to jcc-lmloader@JCC.com