In-Memory Big Data Analysis with Oracle Exalytics

Mark Rittman, Technical Director, Rittman Mead
Oracle Openworld 2012, San Francisco, October 2012
Mark Rittman

• Mark Rittman, Co-Founder of Rittman Mead
• Oracle ACE Director, specialising in Oracle BI&DW
• 14 Years Experience with Oracle Technology
• Regular columnist for Oracle Magazine
• Author of two Oracle Press Oracle BI books
  ‣ Oracle Business Intelligence Developers Guide
  ‣ Oracle Exalytics Revealed
• Writer for Rittman Mead Blog :  
  http://www.rittmanmead.com/blog
• Email : mark.rittman@rittmanmead.com
• Twitter : @markrittman
About Rittman Mead

• Oracle BI and DW platinum partner
• World leading specialist partner for technical excellence, solutions delivery and innovation in Oracle BI
• Approximately 30 consultants worldwide
• All expert in Oracle BI and DW
• UK based
• Offices in US, Europe (Belgium) and India
• Skills in broad range of supporting Oracle tools:
  ‣ OBIEE
  ‣ OBIA
  ‣ ODIEE
  ‣ Essbase, Oracle OLAP
  ‣ GoldenGate
  ‣ Exadata
What is Oracle Exalytics?

• Hardware Element
  ‣ Sun Fire X4470 M2 server
  ‣ 1TB RAM, 40 Cores, 3.6TB HDD

• Software Element
  ‣ OBIEE 11.1.1.6 with Exalytics Enhancements
  ‣ Oracle Essbase 11.1.2 with Exalytics Enhancements
  ‣ Oracle TimesTen 11.2.2.2 for Exalytics
  ‣ New in v1.1: Support for Oracle Endeca Information Discovery, Golden Gate, ODI
  ‣ Runs on 64-bit Oracle Linux (Exalogic distribution)

• OBIEE and Essbase are licensed as Oracle BI Foundation

• Exalytics features can only be used in conjunction with Exalytics hardware
Exalytics as the Exa-Machine for BI

- Runs the BI layer on a high-performance, multi-core, 1TB server
- In-memory cache used to accelerate the BI part of the stack
- If Exadata addresses 80% of the query performance, Exalytics addresses the remaining 20%
  ‣ Consistent response times for queries
  ‣ In-memory caching of aggregates
  ‣ 40 cores for high concurrency
  ‣ Re-engineered BI and OLAP software that assumes 40 cores and 1TB RAM
Enables High-Density Analysis of Big Data

- BI tier is rarely the bottleneck, but it can be if very dense visualizations are used
  - Sparklines, grid of charts etc
- Exalytics’ 40 cores and 1TB RAM make higher density presentation viable
  - Single query sent to the database
  - Exalytics breaks data up to create microcharts
- Also helps support high numbers of concurrent users (100+)
Also Supports Essbase, and Endeca Information Discovery

- In-Memory Essbase for planning, budgeting and sales analysis-style OLAP applications
- Endeca Information Discovery for search/analytic applications against diverse data
Oracle’s Wider Strategy for Business Analytics

• Connect to all of your data, from all your sources,
• Subject it to the full range of possible inquiry
• Package solutions for known problems and fixed sources, and
• Deploy to PCs and mobile devices, on premise or in the cloud
Focus on Engineered Systems for Large-Scale Data Analysis

- Purpose-built engineered systems for storage & analysis of enterprise-scale data sets
- Oracle Exadata for filtering and advanced analytics against large relational data sets
- Oracle Big Data Appliance supporting Oracle NoSQL, R and Apache Hadoop
Big Data: New Datasets of Higher Variety, Volume and Velocity

- Traditional BI datasets have been relatively small, and well structured
- Financial data and other metrics, with attributes and hierarchies to slice-and-dice it
- Big data is all about collecting and analyzing data sets of wider scope
  - Volume - TBs of data collected from sensors, transactions and other low-granularity data sources
  - Variety - unstructured, semi-structured as well as structured sources
  - Velocity - data arriving in real-time, and analyzed in real-time
Big Data in the Context of the Oracle Exalytics Platform

- Exalytics, through in-memory aggregates and InfiniBand connection to Exadata, can analyze vast (structured) datasets held in relational and OLAP databases
- Endeca Information Discovery can analyze unstructured and semi-structured sources
- InfiniBand connector to Big Data Applicance + Hadoop connector in OBIEE supports analysis via Map/Reduce
- Oracle R distribution + Oracle Enterprise R supports SAS-style statistical analysis of large data sets, as part of Oracle Advanced Analytics Option
Oracle Advanced Analytics : In-Database Predictive Analytics

- In-Database Predictive Analytics and Statistical Analysis
- Massively-Scalable, able to analyze huge volumes of data
- Exposed through SQL and R, enabling broad usage

Comprehensive Predictive Analytic Platform Built-Inside the Database
- Data Mining, Text Mining
- Statistical Analysis (built on R)
- Built for Data Scientists/Analysts

Scalable and Parallel
Tightly-Integrated with SQL
Works Inside Exadata and Big Data Appliance
What is R, and Oracle R Enterprise?

- R is a statistical language similar to Base SAS, or SPSS
- Open-source, run by the R Project (http://www.r-project.org)
- R environment is a suite of client/server products for statistical data manipulation and graphical analysis
- Modeling and Analysis performed in-memory using “frames”
- Enhanced by community-contributed packages
- R distribute the open-source version of R with Oracle Linux
- **Oracle R Enterprise** extends R to allow analysis against frames stored in Oracle tables, views and embed R scripts in database PL/SQL packages
Typical R Graphical Output
Capabilities of R Compared to SQL (Built-In Stats Functions)

- R provides a wide variety of statistical and graphical techniques
- Linear and non-linear modeling, classical statistical tests, time-series analysis
- Classification, clustering and other capabilities
- Matrix arithmetic, with scalar, vector, matrices, list and data frame (aka table) structures
- Extensible through community-contributed packages, and interacts with C++, Java etc
- Available for Oracle Database 11gR2 through the Advanced Analytics Option
- Extends the (free) SQL statistical capabilities provided by Oracle Database
  - Ranking, Windowing, Reporting
  - Lag/Lead, First/Last
  - Linear Regression, Inverse Percentile etc
Oracle R Enterprise

- Regular R is constrained by only working with in-memory datasets (frames)
- Data from tables and other database structures has to be loaded into memory
- Oracle R Enterprise (ORE) removes this constraint by allowing frames to reside in DB
- Automatically exploits database parallelism, plus Oracle scalability / resilience
- ORE provides three key areas of functionality
  - Embedded R
  - In-Database Statistics Engine (R extensions for Oracle SQL)
  - Transparency Layer (access RDBMS-based frames)
Typical R / ORE Topology

- Analyst workstation contains Open Source R client tools
- Models created in-memory on workstation
- ORE provides capability to access datasets stored in Oracle RBDMS, transparently
- Database has ORE embedded within it
- ORE provides data for workstation, and can spawn its own R sessions for in-database R analysis
- Enables lights-out R analysis, plus connectivity to Hadoop and Map/Reduce via Oracle R Connector for Hadoop
Oracle Exalytics In-Memory Machine as a Platform for Endeca & R

- Exalytics makes a great platform for Endeca Information Discovery, and R
  - 1TB RAM, 40 CPU cores
  - Infiniband connection to Exadata and Oracle Big Data Appliance
  - Single supplier, single point of support
  - Single dedicated server for analytics
- All products certified on the Exalytics platform
  - OBIEE / Essbase delivered as core software
  - Oracle EID added separately
  - R included in Oracle Linux distribution
- Connectivity to Oracle Database & Oracle BDA
  - Oracle R Enterprise
  - Hadoop, Map/Reduce
  - R connector for Hadoop
Oracle Exalytics In-Memory Machine as a Platform for Endeca & R

- Exalytics server acts as an analysis workstation “on steroids”
- 1TB Ram + 40 cores for multiple R engines
- Infiniband connection to Exadata and ORE
- Endeca Information Discovery uses 40 CPU cores for massively parallel indexing, analysis
- More of Endeca Server datastore in RAM
- OBIEE uses RAM and cores for TimesTen datamart, in-memory data source federation and Presentation Server caching, performance
- All products benefit from high-end server features and fast connectivity to Exadata + BDA
An Example OBIEE / Endeca / R on Exalytics Scenario

- “Airline On-Time Performance and Causes of Flight Delays” dataset
- Provide by Bureau of Transportation Statistics, Research and Innovative Technology Administration, United States Department of Transportation
- Dataset containing 123M rows of non-stop US domestic flight legs
- Source and destination airports, operator, aircraft type
- Type and duration of delay, delay reason
- Freely-available “big data” set
- What can OBIEE, R and EID tell us?
  - OBIEE - dashboard analysis + drilling
  - EID - discovery, and analysis of supporting information describing delays, reasons etc
  - ORE - deep insight into specific questions
Typical Exalytics Analysis Approach

• Endeca Information Discovery for initial data discovery
  ‣ Find out what’s interesting, aggregate and analyze what we find
• OBIEE + Essbase
  ‣ Analyze in a structured way based on questions we’ve just discovered
• Oracle R Enterprise
  ‣ Dig into the dataset and answer the hard, statistical questions
Oracle Endeca Information Discovery Platform Overview

- A suite of tools to capture, load and then analyze all types of data
- Built around the Oracle Endeca Server, a hybrid search/analytic database
- Full ETL capabilities plus web-based dashboard and data exploration tools
- Additional tools can be licenced for content acquisition and text enrichment / sentiment analysis
Oracle Endeca Server: A Hybrid Search/Analytic Database

- Key to these capabilities is the Oracle Endeca Server and its datastores (databases)
- Proprietary database engine focused on **search** and **analytics**
- Data organized as records, made up of attributes stored as key/value pairs
- No over-arching schema, no tables, self-describing attributes
- Every record can have its own unique set of attributes, with the overall data model emerging over time as data is loaded
- Endeca Server hallmarks:
  - Minimal upfront design
  - Support for “jagged” data
  - Administered via web service calls
  - “No data left behind”
  - “Load and Go”
Inside an Endeca Server Datastore

Attributes unique to the sales data mart / warehouse
e.g. Salesperson, Customer, Channel, Response Type

Shared attributes common to all data sources
eg. Time, City, Brand, Key Account

Sales transaction data
from a data mart or data warehouse

Attributes unique to just this data source,
for example Customer Segment, Category, Socio Class

External data from market
research or customer surveys /
demographics

“Jagged” data, where each row may
use different sets of attributes e.g. customer
demographic and behavioral attributes

Forecast data
from Essbase

Supporting data
and comments from
social media, e.g.
Facebook, Twitter

T : +44 (0) 8446 697 995 or (888) 631 1410 (USA)  E : enquiries@rittmanmead.com  W: www.rittmanmead.com
Endeca Server Search and Document Analysis Capabilities

- Search and unstructured data support adds powerful document analysis capabilities
  - Keyword search, boolean search, parametric search, wildcard search, dimension search and dimension filters
  - Dimension precedence rules
  - Numeric range, geospatial, date/time and security filters
  - Spell correction/suggestion, and “do you mean”-type alternative presentation
  - Find similar, and 1 and 2-way synonyms
  - Stemming and lemmatisation
  - Keyword-in-context snippeting
  - Results clustering, relevance ranking, sorting and paging
  - Multi-language support
Oracle EID Integrator and Studio

- Data is loaded into Oracle Endeca Server datastores using Oracle EID Integrator
  - Data Integration (ETL) tool build on open-source CloverETL tool (Eclipse framework)
  - Oracle EID functionality provided through components that call Endeca Server web services
- User Interface created and delivered using Oracle EID Studio, 100% web-based
  - Create dashboards made up of search, navigation and data analysis components
  - Also provides Endeca Server / Studio admin features
Endeca Information Discovery Search/Analytic Dashboard

- Search capabilities of EID help us explore messy and unfamiliar data.
- “McDonel” corrects to “McDonnell”
- Matches to all variants of “McDonnell Douglas”
- Dashboard tells us that Delta Airlines has suffered most delays using McDonnell Douglas aircraft, regardless of variations of spelling.

Any variant of McDonnell Douglas is retrieved.
Discovery Environment Provides Many Ways to Explore Data

“San Fran” provides completion suggestions across all attributes, allowing us to discover many representations of “San Fran” that may be present in the data.

Similarly, with each step of the data exploration, all available records in current filter set are summarised by facets - prompts for every available attribute, populated by only the valid values that will lead to non-empty results sets, allowing users to uncover relationships and patterns in the data using attributes that they may not even be aware existed.

Unlike search engines, EID has a rich, SQL-like language for aggregating and calculations, to populate graphs and visualizations in the dashboard.
Text Analytics through EID Lexical / Parsing / Enrichment Features

- Flexible key/value data model and unstructured text enrichment capabilities of EID allow text analytics to be combined with data discovery and analytics.

Most common issue for older, MD-88 aircraft is “corroded or cracked skin on the fuselage”

Whilst for newer 777, most common issue is “inoperative emergency lights in the cabin”
Analyzing the Flight Delays Dataset using OBIEE

- OBIEE BI Repository created against the flight delays dataset, modeled as two stars
Creating the OBIEE BI Repository

• Repository contains the logical dimensional model for the flight delays dataset
• Logical tables (facts, dimensions) along with hierarchies and aggregation rules
• Subject area for each fact + linked dimensions
• Combined with other models and subject areas of interest
• Use Tools > Utilities > Oracle BI Summary Advisor to start creating TT aggregates
Creating In-Memory Aggs Step 1: Select Facts and Time Scope

- Select facts from the BMM layer to consider for the exercise
  - Can also generate stats from usage tracking data if none exists (upgrade etc)
  - And can load parameters from a file
- Select start and end date for wizard to consider (optional)
Creating In-Memory Aggregates Step 2: Set Min Query Time, Target

- Select minimum elapsed time for queries, to consider for exercise
- Select physical database details for TT database
Creating In-Memory Aggregates Step 3 : Set Script Target Location

- Summary Advisor will generate a script that has to be run through nqcmd.exe
  - Same type of script as generated by Aggregate Persistence Wizard
- Script can be re-run to refresh aggregates periodically
  - Drops and recreates aggregates, no incremental refresh
Creating In-Memory Aggregates Step 4: Define Stopping Criteria

- Select maximum run-time for summary advisor (0=unlimited time)
  - Useful when lots of summary statistics data to process
- Set limit for how little each additional aggregate improves performance
  - Stops lots of little aggregates that barely improve performance being recommended
  - Ideally, set to around 10% for good balance between good recommendations and good dashboard performance
Creating In-Memory Aggregates Step 5: Run + Filter Recommends

- Aggregate recommendations can then be generated
- Filter page lets you select which ones to implement
  - Suggested approach is to add one aggregate at a time, test and implement
  - For smaller, more tested RPDs you can select all if required
Creating In-Memory Aggregates Step 6: Generate Aggregates

- Wizard then creates the script, which can be run through the nqcmd utility
  ‣ Command-line access to the BI Server
- Run command either at the developer workstation, or on the Exalytics server
- Re-run at any time to refresh aggregate tables with new data

```
cd [obiee_client_home]\oraclebi\orahome\bifoundation\server\bin\nqcmd.exe
-d exalytics -u biadmin -p welcome1
-s c:\summary_advisor_scripts\agg_wiz.sql
```
What Does the Summary Advisor Create?

- Aggregate tables in the TimesTen database
  - One per aggregate fact table
  - One per shared dimension
- Physical tables in the Oracle BI Repository physical layer
- Logical Table Sources in the BMM layer
  - All mapped in and active
Final Exalytics Dashboard: “Speed of Thought” Analysis

123m rows of data, analyzed live with detail in Oracle DB, and aggregates in TimesTen

“Go-less” prompts and dashboard controls for instant response to filter changes

Interactive visuals, in the form of maps, graphs, tables, scorecards and KPIs
Using R to Answer In-Depth, Statistical Questions

• Are some airports more prone to delays than others?
• Are some days of the week likely to see fewer delays than others?
  ‣ Are these differences significant?
• How do arrival delay distributions differ for the best and worst 3 airlines compared to the industry?
  ‣ Are there significant differences among airlines?
• For American Airlines, how has the distribution of delays for departures and arrivals evolved over time?
• How do average annual arrival delays compare across select airlines?
  ‣ What is the underlying trend for each airline?
Preparing the Dataset for R, and Running R Queries

• Create R frames using data from Oracle RDBMS, using ORE transparency layer
• Create R queries to manipulate flight delays data
• Build regression models
• Score and rank data
• 40 cores and 1TB RAM in Exalytics allows multiple R engines to be spawned, processing larger datasets than desktop workstation could support

```r
ontimeSubset <- subset(ONTIME_S, UNIQUECARRIER %in% c("AA", "AS", "CO", "DL", "WN", "NW"))
res22 <- with(ontimeSubset, tapply(ARRDELAY, list(UNIQUECARRIER, YEAR), mean, na.rm = TRUE))
g_range <- range(0, res22, na.rm = TRUE)
rindex <- seq_len(nrow(res22))
cindex <- seq_len(ncol(res22))
par(mfrow = c(2,3))
for(i in rindex) {
  temp <- data.frame(index = cindex, avg_delay = res22[i,])
  plot(avg_delay ~ index, data = temp, col = "black",
       axes = FALSE, ylim = g_range, xlab = "", ylab = "",
       main = attr(res22, "dimnames")[[1]][i])
  axis(1, at = cindex, labels = attr(res22, "dimnames")[[2]])
  axis(2, at = 0:ceiling(g_range[2]))
  abline(lm(avg_delay ~ index, data = temp), col = "green")
  lines(lowess(temp$index, temp$avg_delay), col="red")
}
```

Saturday, 29 September 12
Integrating with OBIEE and Oracle BI Publisher

• R scripts can be embedded in BI Publisher data models
• Results returned as image vectors in XML, and rendered as BI Publisher output
• R scripts can also be referenced in functions etc and included in OBIEE RPD
R Analysis Output within the OBIEE Dashboard

Display flight delay per airport for top N busiest airports with parameters that are passed to live R engines, using R script in BIP data model.

Regression analysis used to predict average delay for a route, using ORE integration within OBIEE BI Repository.
Summary

• Analytics options available with Oracle Database and Oracle Fusion Middleware support a wide range of analytic tools and engines
• Oracle Exalytics is an excellent platform to run these on, based on RAM and CPU #
• OBIEE, with TimesTen for Exalytics and the Summary Advisor, supports “speed of thought” analytics using a rich, interactive dashboard
• Endeca Information Discovery provides a search / analytic interface to enable you to discover the questions that need answering
• Oracle R Enterprise, part of the Advanced Analytics Option for Oracle Database, enables deep analysis and insights using the R statistical language and environment
• A combined, integrated analysis toolset based on an Oracle Engineered System
More Information

• Thank you for attending this presentation
• More information can be found at http://www.rittmanmead.com
• Contact us at info@rittmanmead.com or mark.rittman@rittmanmead.com
• Look out for our book, “Oracle Business Intelligence Developers Guide” out now!
• Follow-us on Twitter (@rittmanmead) or Facebook (facebook.com/rittmanmead)