Oracle Big Data Spatial & Graph Social Media Analysis - Case Study

Mark Rittman, CTO, Rittman Mead
BIWA Summit 2016, San Francisco, January 2016
About Rittman Mead

• Oracle Gold Partner with offices in the UK and USA (Atlanta)
• 70+ staff delivering Oracle BI, DW, Big Data and Advanced Analytics projects
• Oracle ACE Director (Mark Rittman, CTO) + 2 Oracle ACEs
• Significant web presence with the Rittman Mead Blog (http://www.rittmanmead.com)
• Regular users of social media (Facebook, Twitter, Slideshare etc)
• Regular column in Oracle Magazine and other publications
• Hadoop R&D lab for “dogfooding” solutions developed for customers
Business Scenario

• Rittman Mead want to understand drivers and audience for their website
  ‣ What is our most popular content? Who are the most in-demand blog authors?
  ‣ Who are the influencers? What communities exist around our web presence?
• Three data sources in scope:
  - RM Website Logs
  - Twitter Stream
  - Website Posts, Comments etc
OBIEE and BDD for the “What” and “Why” Questions…

- Counts of page views, tweets, mentions etc helped us understand **what** content was popular
- Analysis of tweet sentiment, meaning and correlation with content answered **why**

Combine with Oracle Big Data SQL for structured OBIEE dashboard analysis

- What pages are people visiting?
- Who is referring to us on Twitter?
- What content has the most reach?

Combine with site content, semantics, text enrichment Catalog and explore using Oracle Big Data Discovery

- Why is some content more popular?
- Does sentiment affect viewership?
- What content is popular, where?
But Who Are The Influencers In Our Community?

• Previous counts assumed that all tweet references equally important
• But some Twitter users are far more influential than others
  ‣ Sit at the centre of a community, have 1000’s of followers
  ‣ A reference by them has massive impact on page views
  ‣ Positive or negative comments from them drive perception
• Can we identify them?
  ‣ Potentially “reach out” with analyst program
  ‣ Study what website posts go “viral”
  ‣ Understand out audience, and the conversation, better
What Communities and Networks Are Our Audience?

- Rittman Mead website features many types of content
  - Blogs on BI, data integration, big data, data warehousing
  - Op-Eds (“OBIEE12c - Three Months In, What’s the Verdict?”)
  - Articles on a theme, e.g. performance tuning
  - Details of new courses, new promotions
- Different communities likely to form around these content types
- Different influencers and patterns of recommendation, discovery
- Can we identify some of the communities, segment our audience?

Identify group of people that are close to each other – e.g. target group marketing
Graph Example: RM Blog Post Referenced on Twitter

Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools

There comes the point in any sufficiently complex or difficult problem diagnosis that the log files in OBIEE alone are not sufficient for building up a complete picture of what’s going on. Even with the debug/trace data that Presentation Services and other components can be configured precisely to write you’re sometimes just left having to guess what is going on inside the black box of each of the OBIEE system components.

Here we’re going to look at a couple of examples of lifting the lid just a little bit further on what OBIEE is up to, using standard Linux diagnostic tools. These are not something to be reaching for in the first instance, but more getting on to a last resort. Almost always...
Network Effect Magnified by Extent of Social Graph

Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools

There comes the point in any sufficiently complex or difficult problem diagnosis that the log files in OBIEE alone are not sufficient for building up a complete picture of what’s going on. Even with the debug/trace data that Presentation Services and other components can be configured precisely to write you’re sometimes just left having to guess what is going on inside the black box of each of the OBIEE system components.

Here we’re going to look at a couple of examples of lifting the lid just a little bit further on what OBIEE is up to, using standard Linux diagnostic tools. These are not something to be reaching for in the first instance, but more getting on to a last resort. Almost always
Retweets by Influential Twitter Users Drive Visits

RT: Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools http://t.co/gFcUPOm5pI

Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools http://t.co/gFcUPOm5pI
Retweets, Mentions and Replies Create Communities

#bigdatasql

#thatshesaid
Property Graph Terminology

- Node, or “Vertex”
- Directed Connection, or “Edge”
- Mentions
- Retweets
Determining Influencers - Factors to Consider

- Different types of Twitter interaction could imply more or less “influence”

  - **Retweet** of another user’s Tweet implies that person is worth quoting or you endorse their opinion

  - **Reply to** another user’s tweet could be a weaker recognition of that person’s opinion or view

  - **Mention** of a user in a tweet is a weaker recognition that they are part of a community / debate
Relative Importance of Edge Types Added via Weights

- **Edge Property**
  - Retweet, Weight = 100

- **Edge Property**
  - Mentions, Weight = 30

- **Gwen (Chen) Shapira**
  - TWEETS: 14,900
  - FOLLOWING: 2,012
  - FOLLOWERS: 5,444
  - LIKES: 1,797
  - Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools [http://t.co/gFcUP0m5pI](http://t.co/gFcUP0m5pI)

- **Neselovskyi, Borys**
  - TWEETS: 552
  - FOLLOWING: 158
  - FOLLOWERS: 76
  - Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools [http://t.co/gFcUP0m5pI](http://t.co/gFcUP0m5pI)
ODI12c used to prepare two files in Oracle Flat File Format
• Extracted vertices and edges from existing data in Hive
  ‣ Wrote vertices (Twitter users) to .opv file, edges (RTs, replies etc) to .ope file
• For exercise, only considered 2-3 days of tweets
  ‣ Did not include follows (user A followed user B) as not reported by Twitter Streaming API
  ‣ Could approximate larger follower networks through multiplying weight of edge by follower scale
    - Useful for Page Rank, but does it skew actual detection of influencers in exercise?
### Oracle Flat File Format Vertices and Edge Files

**Vertex File (.opv)**

- Unique ID for the vertex
- Property name ("name")
- Property value datatype (1 = String)
- Property value ("markrittman")

**Edge File (.ope)**

- Unique ID for the edge
- Leading edge vertex ID
- Trailing edge vertex ID
- Edge Type ("mentions")
- Edge Property ("weight")
- Edge Property datatype and value
Loading Edges and Vertices into HBase

```java
cfg = GraphConfigBuilder.forPropertyGraphHbase()
    .setName("connectionsHBase")
    .setZkQuorum("bigdatalite").setZkClientPort(2181)
    .setZkSessionTimeout(120000).setInitialEdgeNumRegions(3)
    .setInitialVertexNumRegions(3).setSplitsPerRegion(1)
    .addEdgeProperty("weight", PropertyType.DOUBLE, "1000000")
    .build();

opg = OraclePropertyGraph.getInstance(cfg);
opg.clearRepository();

vfile="../../data/biwa_connections.opv"
efile="../../data/biwa_connections.ope"
opgdl=OraclePropertyGraphDataLoader.getInstance();
opgdl.loadData(opg, vfile, efile, 2);

// read through the vertices
opg.getVertices();

// read through the edges
opg.getEdges();
```

Uses “Gremlin” Shell for HBase
- Creates connection to HBase
- Sets initial configuration for database
- Builds the database ready for load

- Defines location of Vertex and Edge files
- Creates instance of OraclePropertyGraphDataLoader
- Loads data from files

- Prepares the property graph for use
- Loads in Edges and Vertices
- Now ready for in-memory processing
Calculating Most Influential Tweeters Using Page Rank

vOutput="/tmp/mygraph.opv"
eOutput="/tmp/mygraph.ope"
OraclePropertyGraphUtils.exportFlatFiles(opg, vOutput, eOutput, 2, false);

session = Pgx.createSession("session-id-1");
analyst = session.createAnalyst();
graph = session.readGraphWithProperties(opg.getConfig());
rank = analyst.pagerank(graph, 0.001, 0.85, 100);
rank.getTopKValues(5);

• Initiates an in-memory analytics session
• Runs Page Rank algorithm to determine influencers
• Outputs top ten vertices (users)

Top 10 vertices

--- PgxVertex with ID 1=0.13885623487462861
--- PgxVertex with ID 3=0.08686102641801993
--- PgxVertex with ID 101=0.06757752513733056
--- PgxVertex with ID 6=0.06743774001139484
--- PgxVertex with ID 37=0.0481517609757462
--- PgxVertex with ID 17=0.042234536894569276
--- PgxVertex with ID 29=0.04109794527311113
--- PgxVertex with ID 65=0.032058649698044187
--- PgxVertex with ID 15=0.023075360575195276
--- PgxVertex with ID 93=0.019265959946506813
Calculating Most Influential Tweeters Using Page Rank

v1=opg.getVertex(11); v2=opg.getVertex(31); v3=opg.getVertex(1011); \
 v4=opg.getVertex(61); v5=opg.getVertex(371); v6=opg.getVertex(171); \
 v7=opg.getVertex(291); v8=opg.getVertex(651); v9=opg.getVertex(151); \
 v10=opg.getVertex(931);
System.out.println("Top 10 influencers: \n " + v1.getProperty("name") + \
  "\n " + v2.getProperty("name") + \
  "\n " + v3.getProperty("name") + \
  "\n " + v4.getProperty("name") + \
  "\n " + v5.getProperty("name") + \
  "\n " + v6.getProperty("name") + \
  "\n " + v7.getProperty("name") + \
  "\n " + v8.getProperty("name") + \
  "\n " + v9.getProperty("name") + \
  "\n " + v10.getProperty("name"));

Note:
Over a 3-day period in May 2015
Twitter users referencing RM website + staff accounts

Top 10 influencers:
markrittman
rmoff
rittmanmead
mRainey
JeromeFr
Nephentur
borkur
BIExperte
i_m_dave
dw_pete
Visualising Property Graphs with Cytoscape

- Open source graph analysis tool with Oracle Big Data Graph and Spatial Plug-in
- Available shortly from Oracle, connects to Oracle NoSQL or HBase and runs Page Rank etc
- Alternative to command-line for In-Memory Analytics once base graph created
Calculating Top 10 Users using Page Rank Algorithm

Top 10 influencers:
- markrittman
- rmoff
- rittmanmead
- mRainey
- JeromeFr
- Nephentur
- borkur
- BIExperte
- i_m_dave
- dw_pete
Visualising the Social Graph Around Particular Users
Calculating Shortest Path Between Users
Edge Bundling to Better Illustrate Connection Frequency
Determining Communities via Twitter Interactions
Determining Communities via Twitter Interactions

- Clusters based on actual interaction patterns, not hashtags
- Detects real communities, not ones that exist just in-theory
Conclusions, and Further Reading

- Tools such as OBIEE are great for understanding what (counts, page views, popular items)
- Oracle Big Data Discovery can be useful for understanding “why?” (sentiment, terms etc)
- Graph Analysis can help answer “who”?
  - Who are our audience? What are our communities? Who are their important influencers?
- Oracle Big Data Graph and Spatial can answer these questions to “big data” scale
- Articles on the Rittman Mead Blog
  - http://www.rittmanmead.com/category/oracle-big-data-
  - http://www.rittmanmead.com/category/oracle-big-data-
- Rittman Mead offer consulting, training and managed services
  - http://www.rittmanmead.com/bigdata
Oracle Big Data Spatial & Graph Social Media Analysis - Case Study

Mark Rittman, CTO, Rittman Mead
BIWA Summit 2016, San Francisco, January 2016