

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

Mark Rittman, CTO, Rittman Mead
OTN EMEA Tour, May 2016



OTN EMEA
tour 2016

About the Speaker

- 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 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:

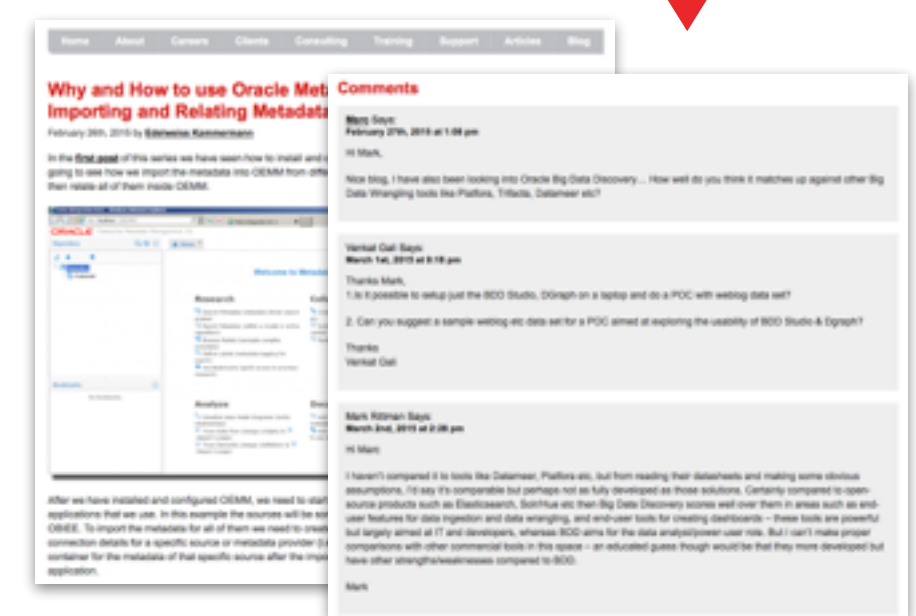
RM Website Logs

```
202.46.48.28 -- [23/Feb/2014:01:03:37 +0000] "GET HTTP/1.1" 200 40440 "-" Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0)
119.63.193.195 -- [23/Feb/2014:01:04:57 +0000] "GET / HTTP/1.1" 200 40440 "-" Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.0)
66.249.74.116 -- [23/Feb/2014:01:26:31 +0000] "GET /2011/03/update-on-the-oracle-press-book/ HTTP/1.1" 200 11585 "-" Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
66.249.74.116 -- [23/Feb/2014:01:52:48 +0000] "GET /blog/page/336/ HTTP/1.1" 200 10600 "-" Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)
100.76.5.222 -- [23/Feb/2014:01:56:10 +0000] "GET /2010/09/rittman-mead-at-oracle-openworld-2010-san-francisco/ HTTP/1.1" 304 "-" Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)
100.76.5.166 -- [23/Feb/2014:02:33:53 +0000] "GET /2009/05/dynamic-sql/ HTTP/1.1" 200 11978 "-" Mozilla/5.0 (compatible; Baiduspider/2.0; +http://www.baidu.com/search/spider.html)
91.232.96.2 -- [23/Feb/2014:02:37:19 +0000] "GET /blog/page/33/ HTTP/1.1" 200 95831 "-" Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
91.232.96.2 -- [23/Feb/2014:02:38:01 +0000] "GET /2012/05/hfm-11-1-2-2-%E2%80%93-new-features-part-1/ HTTP/1.1" 200 50539 "-" Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
91.232.96.2 -- [23/Feb/2014:02:38:09 +0000] "POST /wp-comments-post.php HTTP/1.1" 302 - "http://31.221.34.123/2012/05/hfm-11-1-2-2-%E2%80%93-new-features-part-1/" Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1)"
91.232.96.2 -- [23/Feb/2014:02:38:10 +0000] "GET /2012/05/
```

Twitter Stream

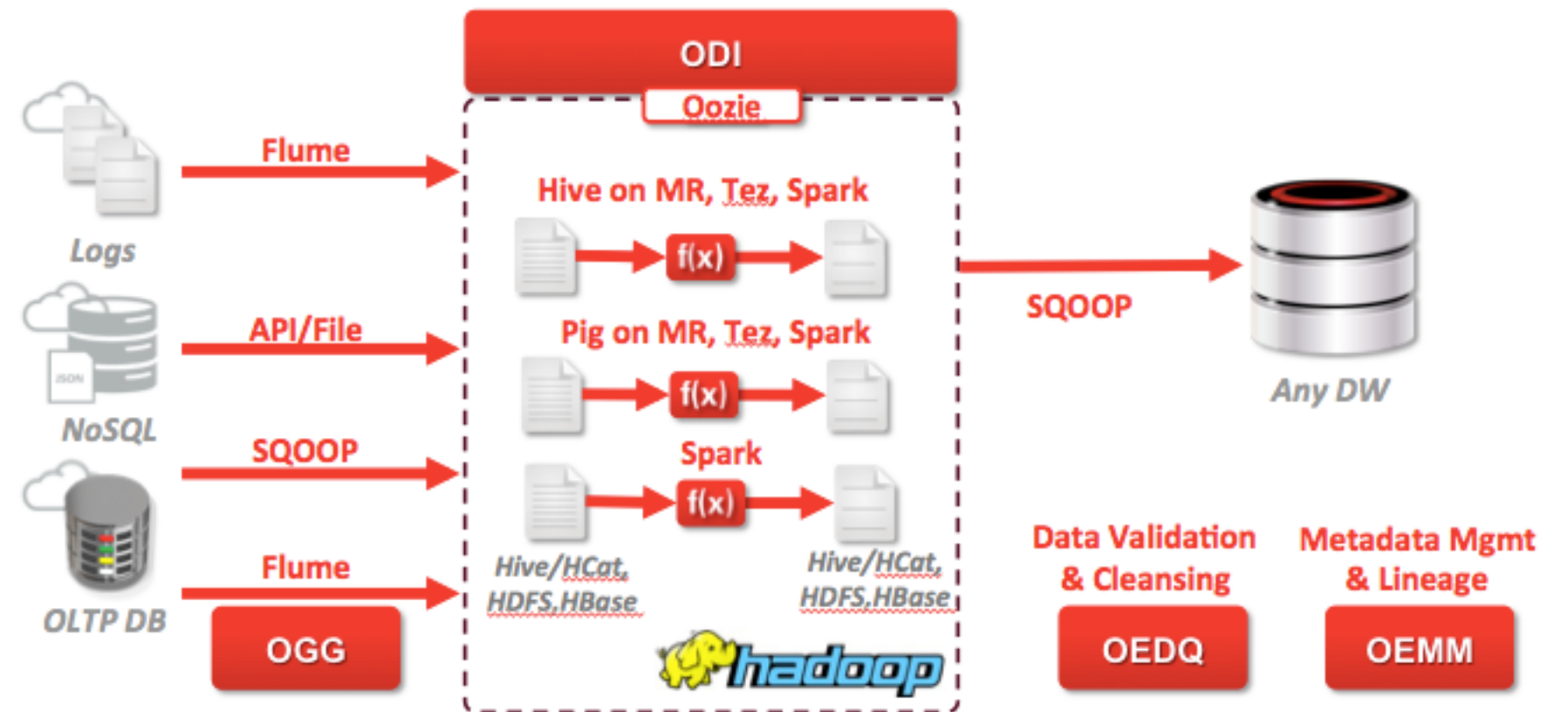


Website Posts, Comments etc



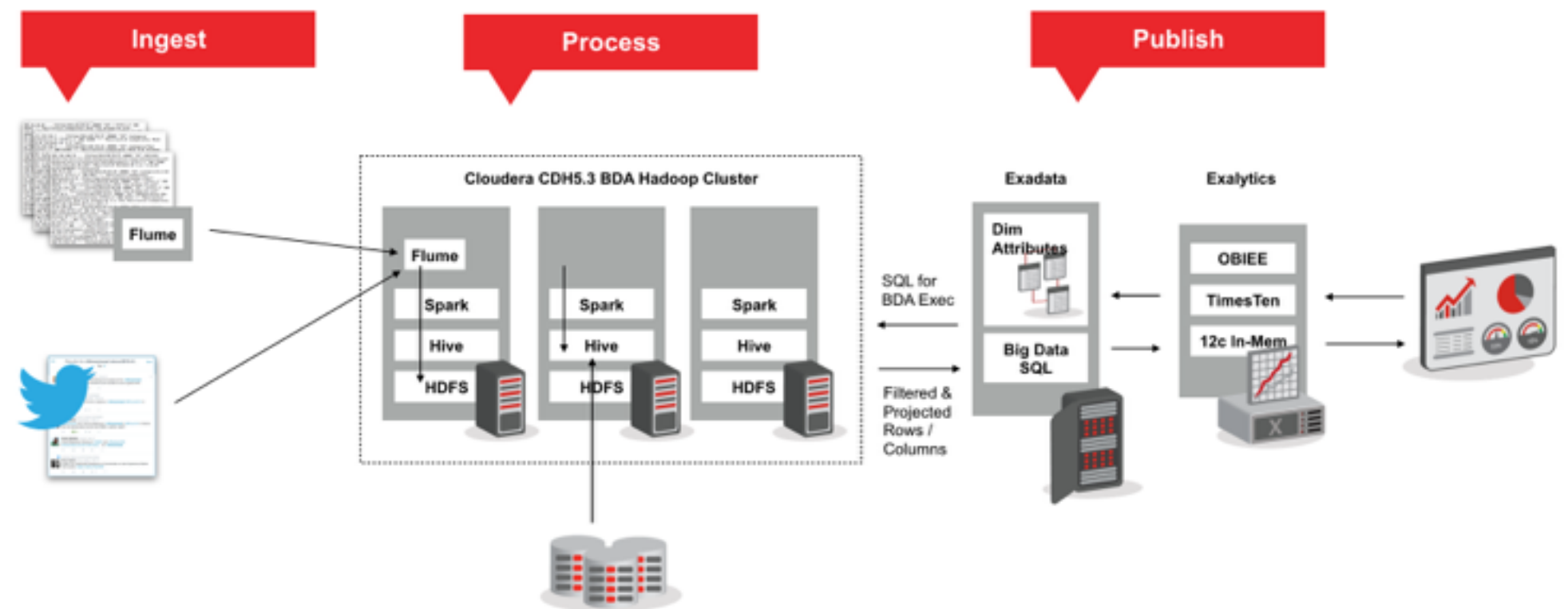
Real-Time & Batch Log & Event Ingestion : ODI12c

- ODI provides an excellent framework for running Hadoop ETL jobs
 - ▶ ELT approach pushes transformations down to Hadoop - leveraging power of cluster
- Hive, HBase, Sqoop and OLH/ODCH KMs provide native Hadoop loading / transformation
 - ▶ Whilst still preserving RDBMS push-down
 - ▶ Extensible to cover Pig, Spark etc
- Process orchestration
- Data quality / error handling
- Metadata and model-driven
- New in 12.1.3.0.1 - ability to generate Pig and Spark jobs too



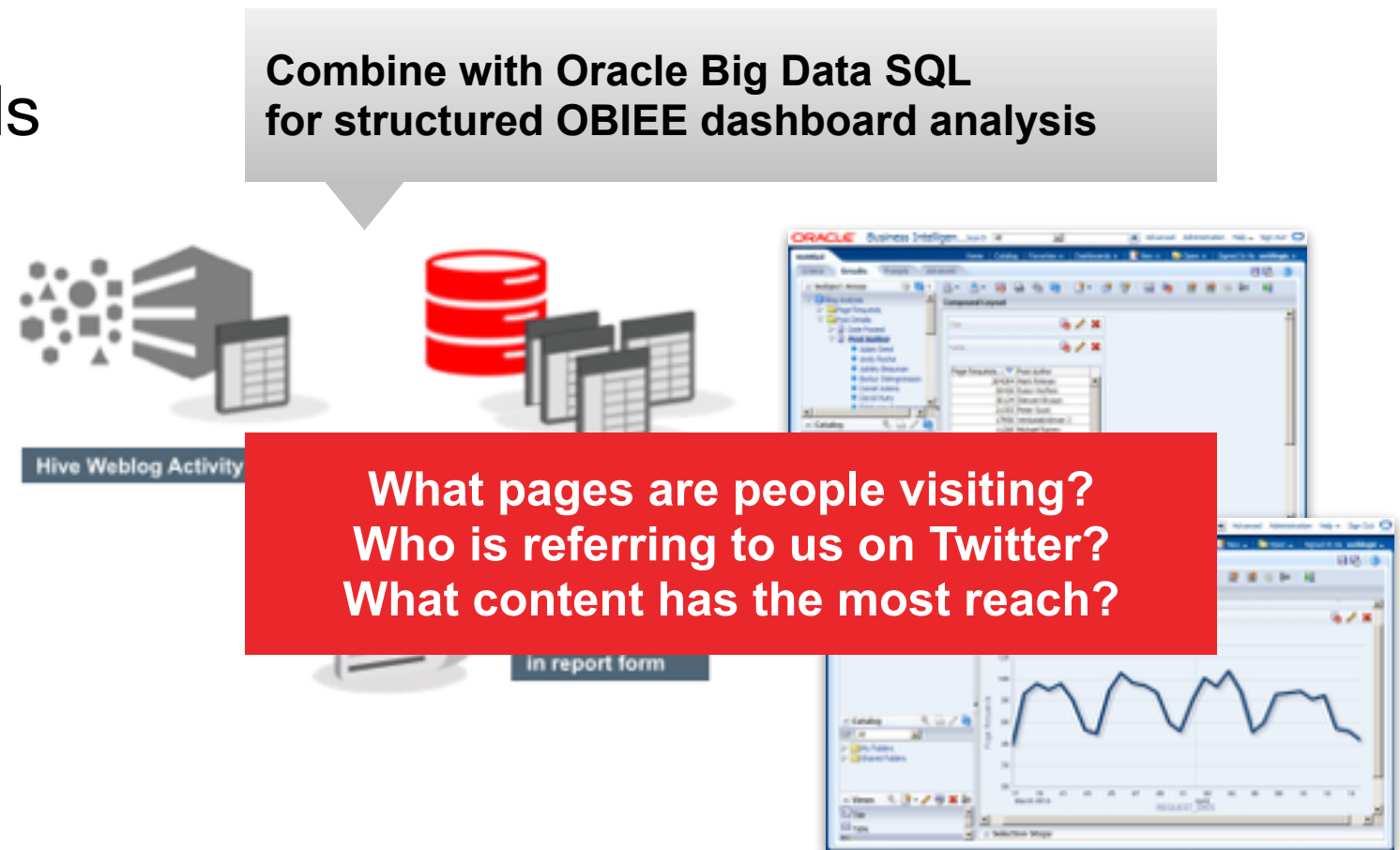
Overall Project Architecture - Phase 1

- Initial iteration of project focused on capturing and ingesting web + social media activity
- Apache Flume used for capturing website hits, page views
- Twitter Streaming API used to capture tweets referring to RM website or RM staff
- Activity landed into Hadoop (HDFS), processed and enriched and presented using Hive



Real-Time Metrics around Site Activity - “What?”

- Provided real-time counts of page views, correlated with Twitter activity stored in Hive tables
- Accessed using Oracle Big Data SQL + joined to Oracle RDBMS reference data
- Delivered using OBIEE reports and dashboards
- Data Warehousing, but cheaper + real-time
- Answered questions such as
 - ▶ What are our most popular site pages?
 - ▶ Which pages attracted the most attention on Twitter, Facebook?
 - ▶ What topics are popular?



Oracle BDD for Data Wrangling + Data Enrichment

- Oracle Big Data Discovery used to go back to the raw event data add more meaning
- Enrich data, extract nouns + terms, add reference data from file, RDBMS etc
- Understand sentiment + meaning of tweets, link disparate + loosely coupled events
- Faceted search dashboards



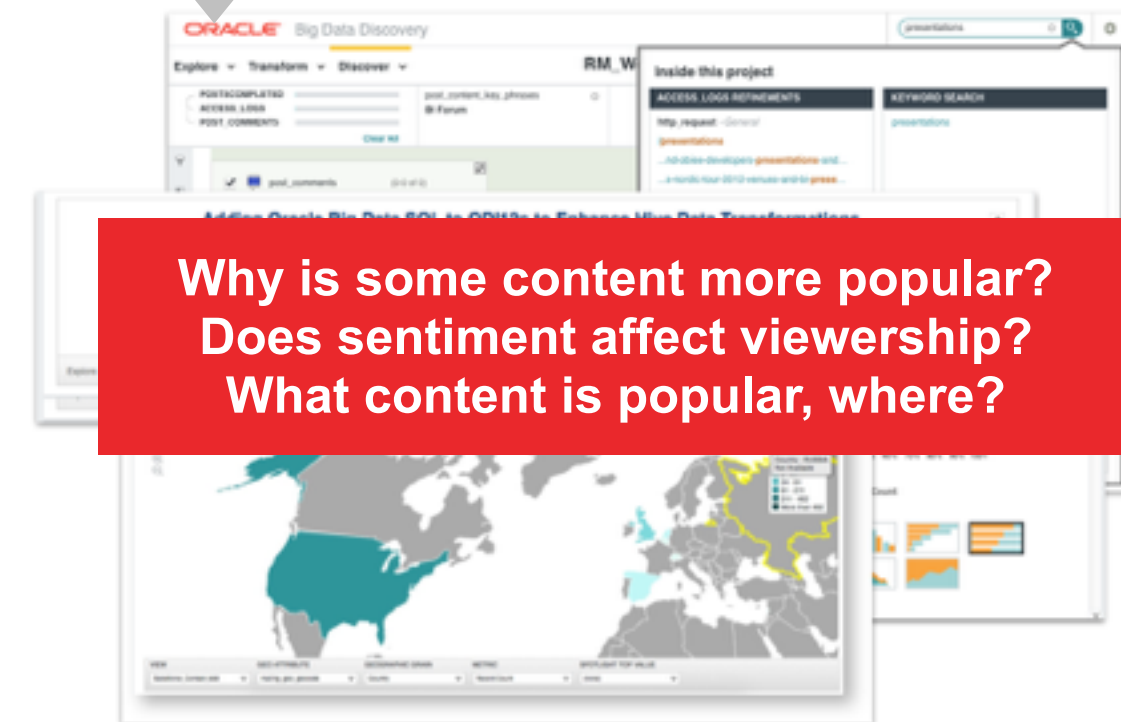
A screenshot of the Oracle Big Data Discovery interface. The interface is divided into several panels. On the left, there's a 'Transformation Editor' showing a SQL query: '1 extractNounGroups(post_title)'. Below it, a table displays data with columns for 'post_id', 'post_title', and 'post_comments'. The main area shows a dashboard with a title 'Adding Oracle Analytics with Kibana and Elasticsearch' and a map of Europe. A sidebar on the right contains 'Inside this project' and 'Keyword Search' sections. A yellow callout box is overlaid on the bottom right of the screenshot, containing the text: 'Real-time ETL Cha', 'Real-time ETL Cha', 'Apps Implementati', 'Apps 11g', and 'Apps Implementation Part, introduction, the Project Life Cycle'.

Answered 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

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

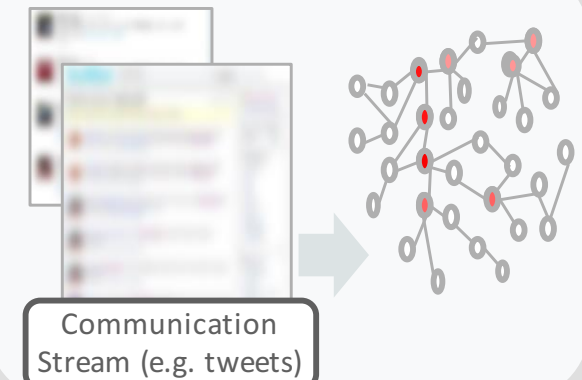


But Who Are The Influencers In Our Community?

- 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

Find out people that are *central* in the given network – e.g. influencer marketing

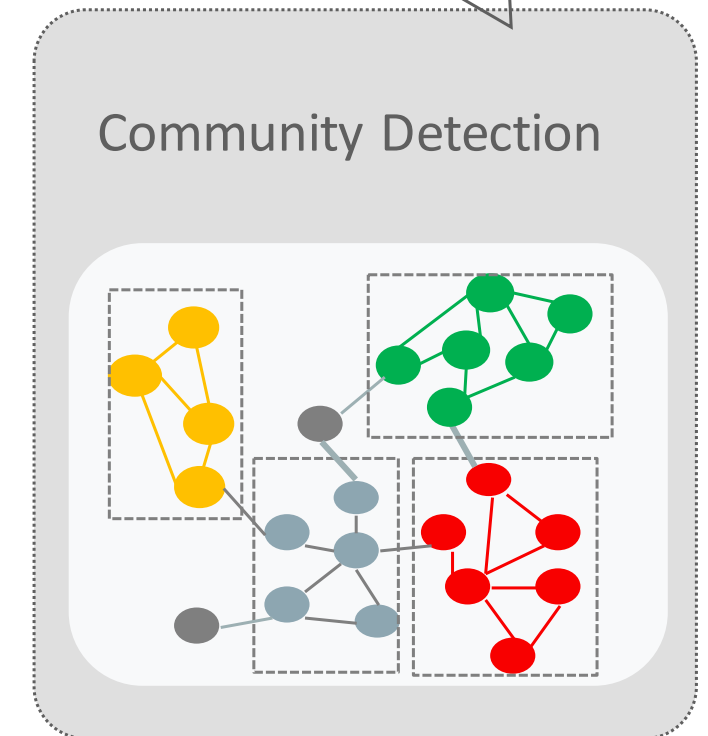
Influencer Identification



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



Tabular (SQL) Query Tools Aimed at Counts + Aggs

- Answers from **Aggregation**


- Who spends the most?
- Who buys the highest margin goods?
- Who is most consistently a top contributor?

- Answers from **Connectivity**

- Who's most influential?
- Which supplier do I depend on the most?
- What is the right product mix for millennials?

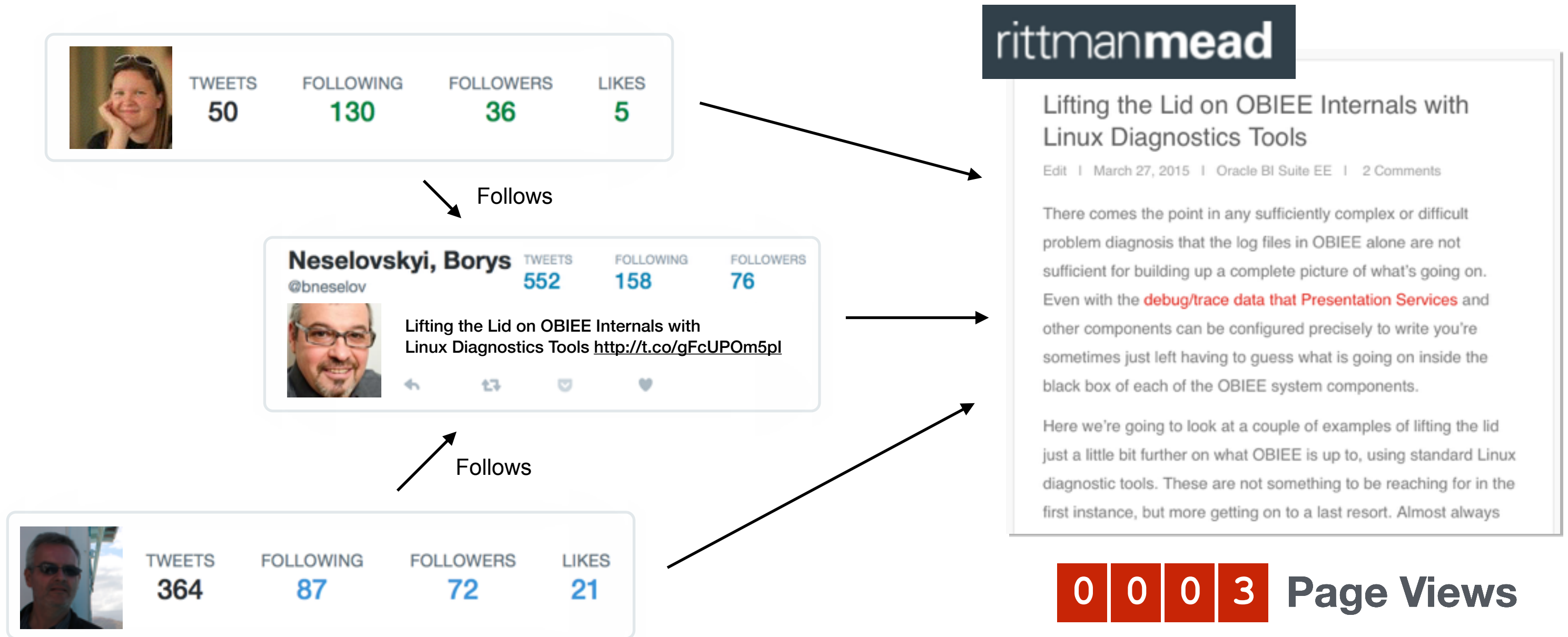


Tabular questions:
Well-suited to SQL-like tools

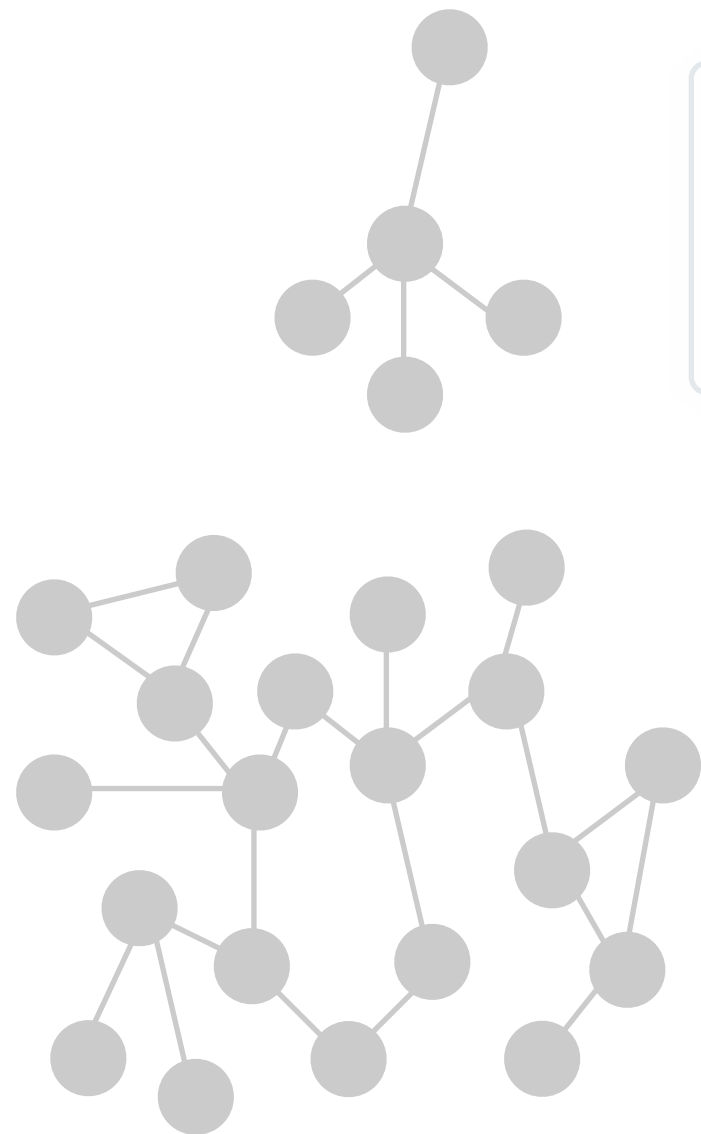


Graph questions:
We need something different!

Graph Example : RM Blog Post Referenced on Twitter




Network Effect Magnified by Extent of Social Graph



Neselovskyi, Borys TWEETS 552 FOLLOWING 158 FOLLOWERS 76
@bneselov

 Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools <http://t.co/gFcUPOm5pl>

Gwen (Chen) Shapira TWEETS 14.9K FOLLOWING 2,012 FOLLOWERS 5,444 LIKES 1,797
@gwenshap FOLLOWS YOU

 Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools <http://t.co/gFcUPOm5pl>

rittmanmead

Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools

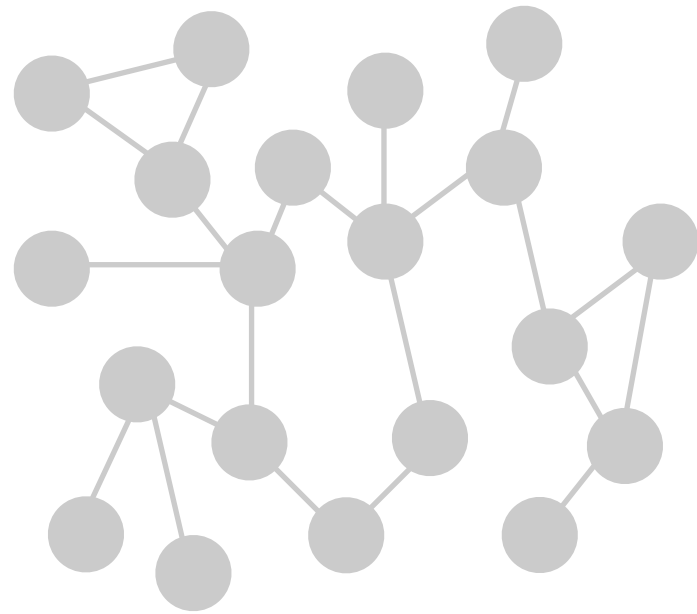
Edit | March 27, 2015 | Oracle BI Suite EE | 2 Comments

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

0 | 0 | 5 | 7 Page Views

Retweets by Influential Twitter Users Drive Visits



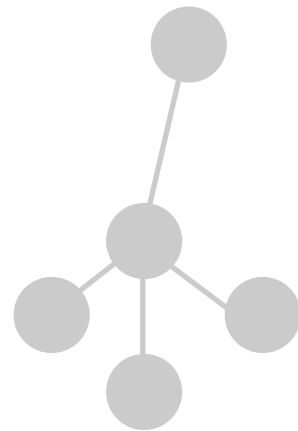
Gwen (Chen) Shapira TWEETS 14.9K FOLLOWING 2,012 FOLLOWERS 5,444 LIKES 1,797
@gwenshap FOLLOWS YOU

 RT: Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools <http://t.co/gFcUPOm5pl>

← ↻ 📌 ❤️

0 0 3 5 Page Views

Retweet



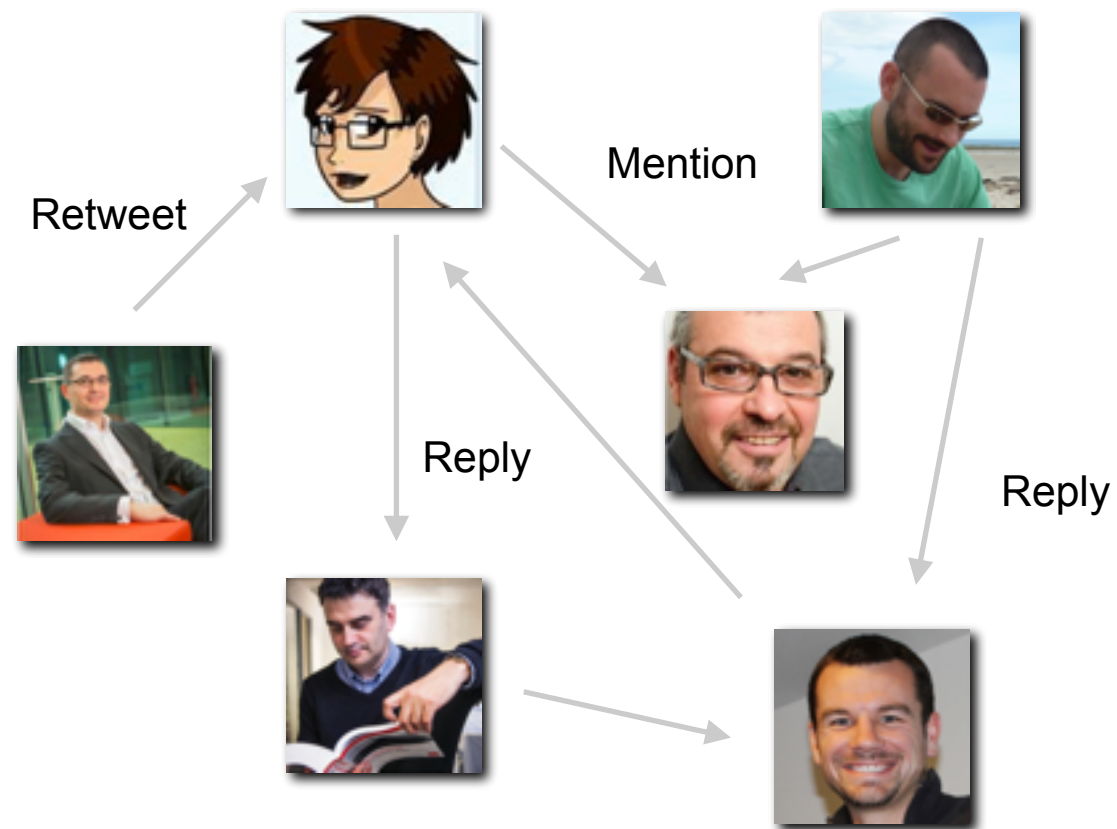
Neselovskyi, Borys TWEETS 552 FOLLOWING 158 FOLLOWERS 76
@bneselov

 Lifting the Lid on OBIEE Internals with Linux Diagnostics Tools <http://t.co/gFcUPOm5pl>

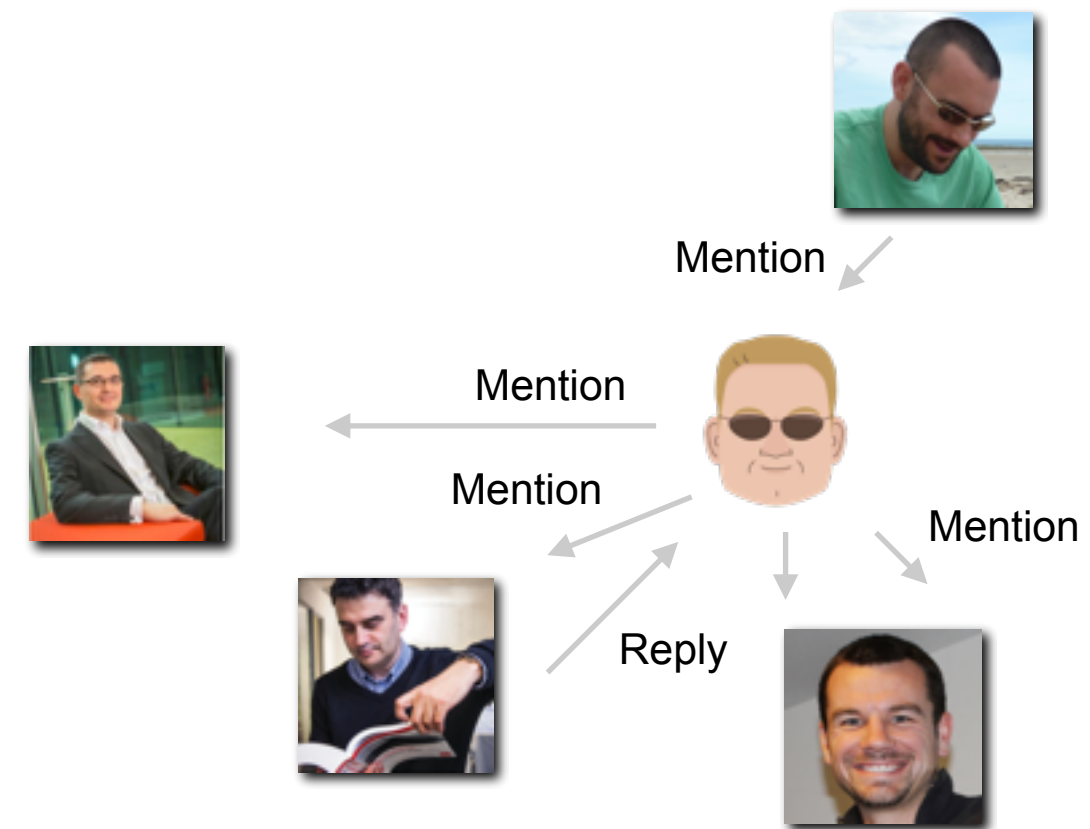
← ↻ 📌 ❤️

0 0 0 3 Page Views

Retweets, Mentions and Replies Create Communities

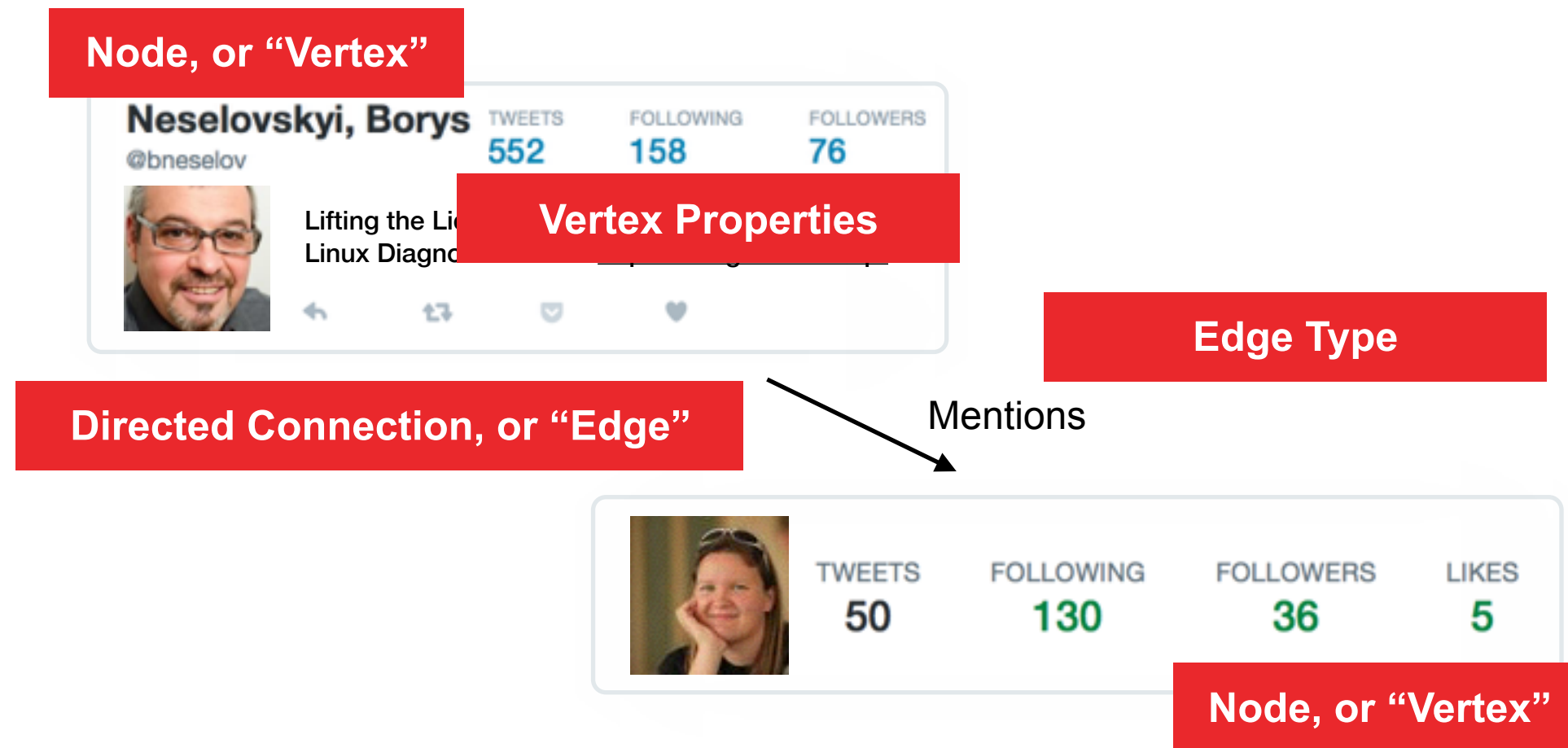


#bigdatasql

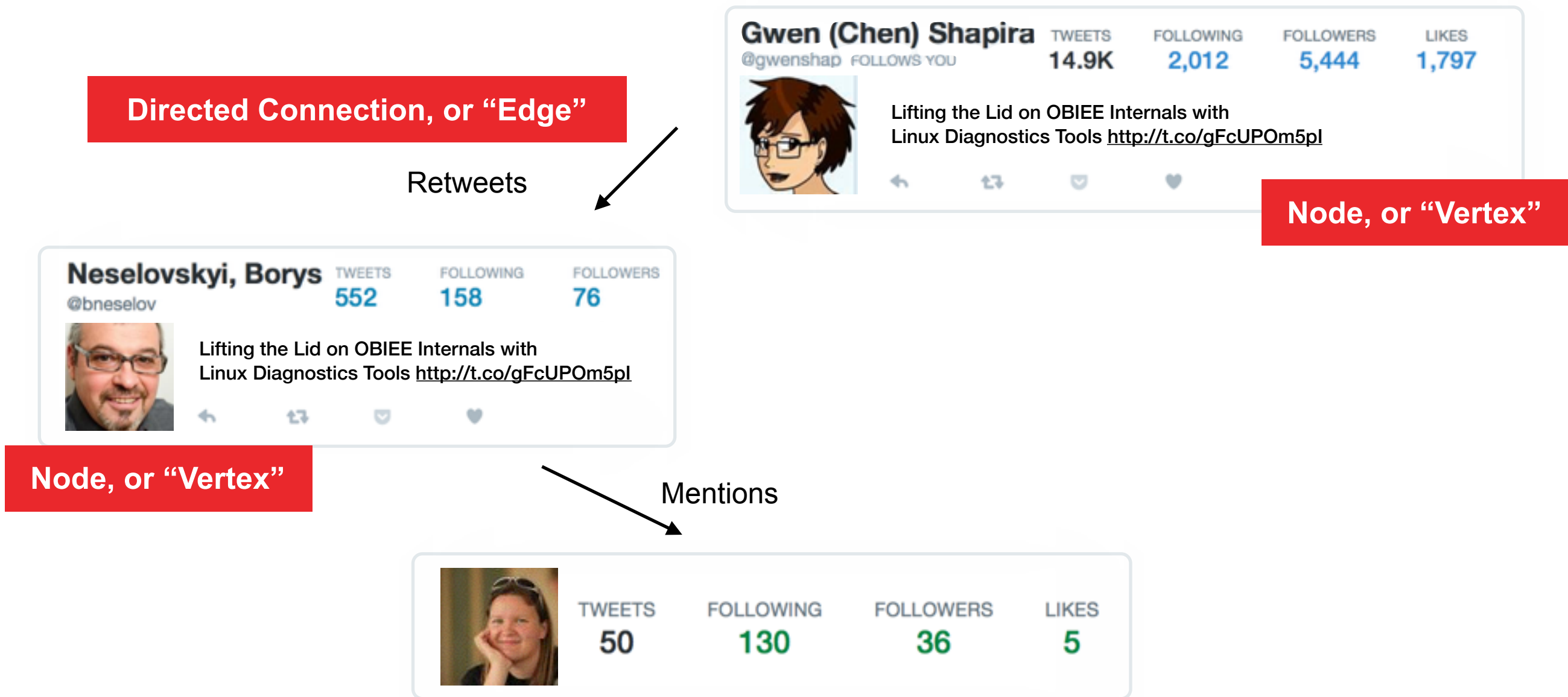


#thatwhatshesaid

Property Graph Terminology



Property Graph Terminology



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



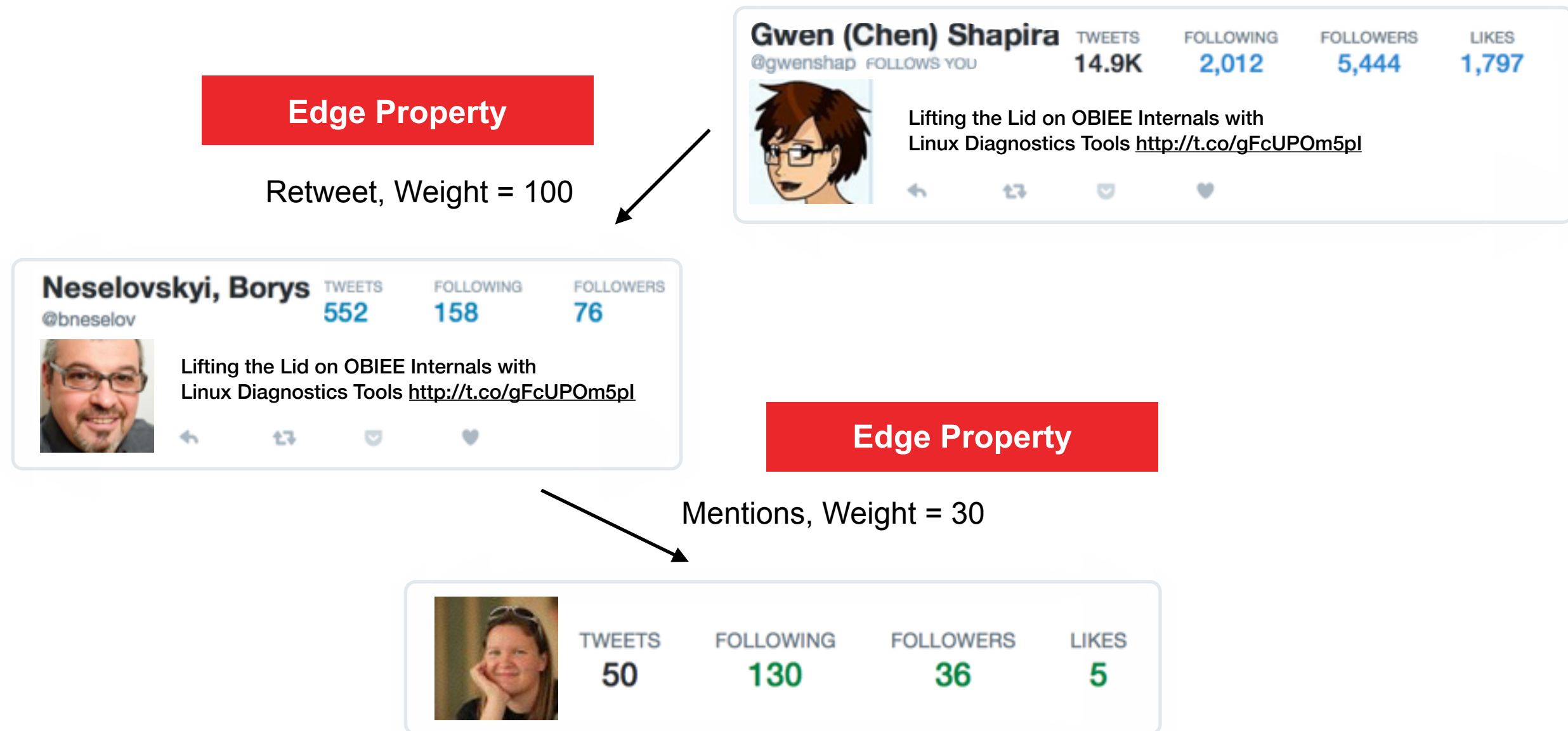
- ▶ **Reply to** another user’s tweet could be a weaker recognition of that person’s opinion or view



- ▶ **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



Oracle Big Data Spatial & Graph

- Graph, spatial and raster data processing for big data
 - ▶ Primarily documented + tested against Oracle BDA
 - ▶ Installable on commodity cluster using CDH
- Data stored in Apache HBase or Oracle NoSQL DB
 - ▶ Complements Spatial & Graph in Oracle Database
 - ▶ Designed for trillions of nodes, edges etc
- Out-of-the-box spatial enrichment services
- Over 35 of most popular graph analysis functions
 - ▶ Graph traversal, recommendations
 - ▶ Finding communities and influencers,
 - ▶ Pattern matching



Oracle Big Data Graph and Spatial Architecture

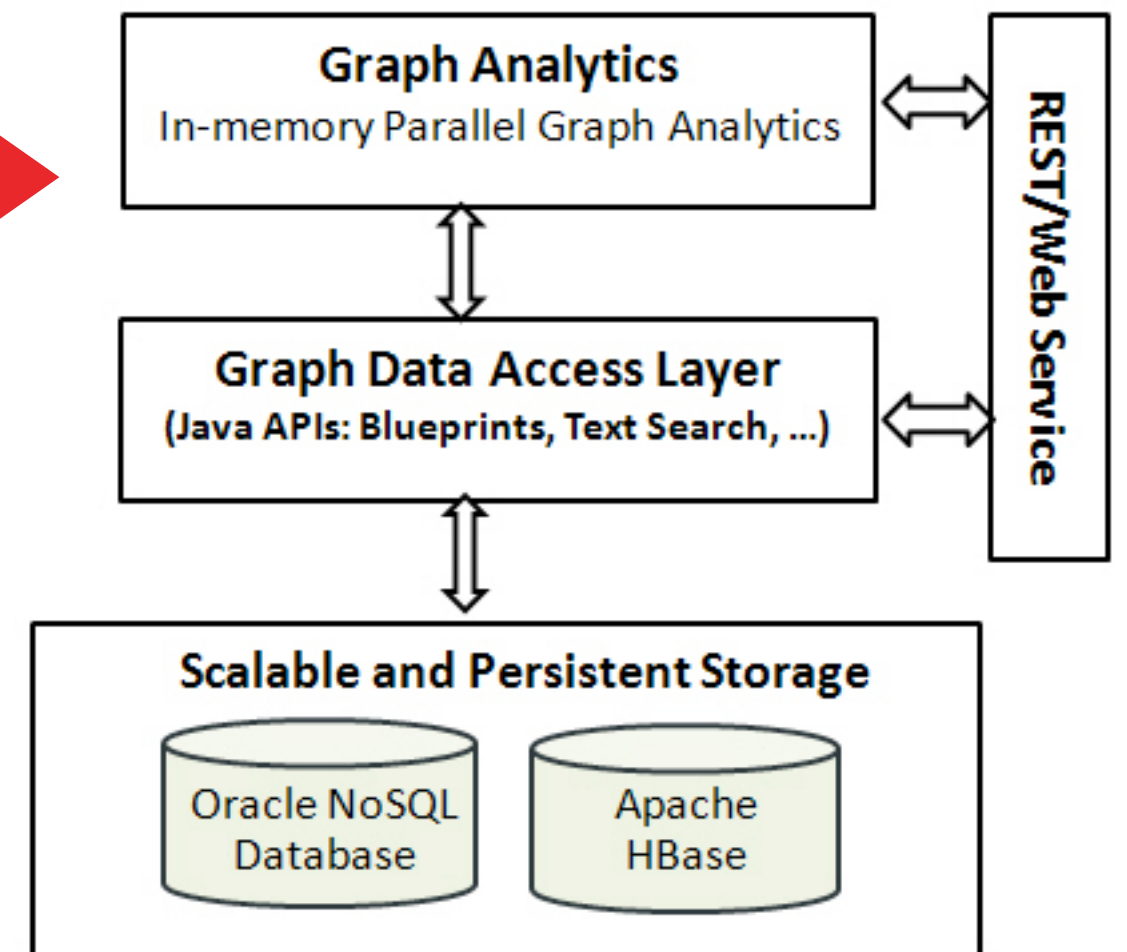
- Data loaded from files or through Java API into HBase
- In-Memory Analytics layer runs common graph and spatial algorithms on data
- Visualised using R or other graphics packaged

Lightning-Fast In-Memory Analytics

- YARN Container
- Standalone Server
- Embedded

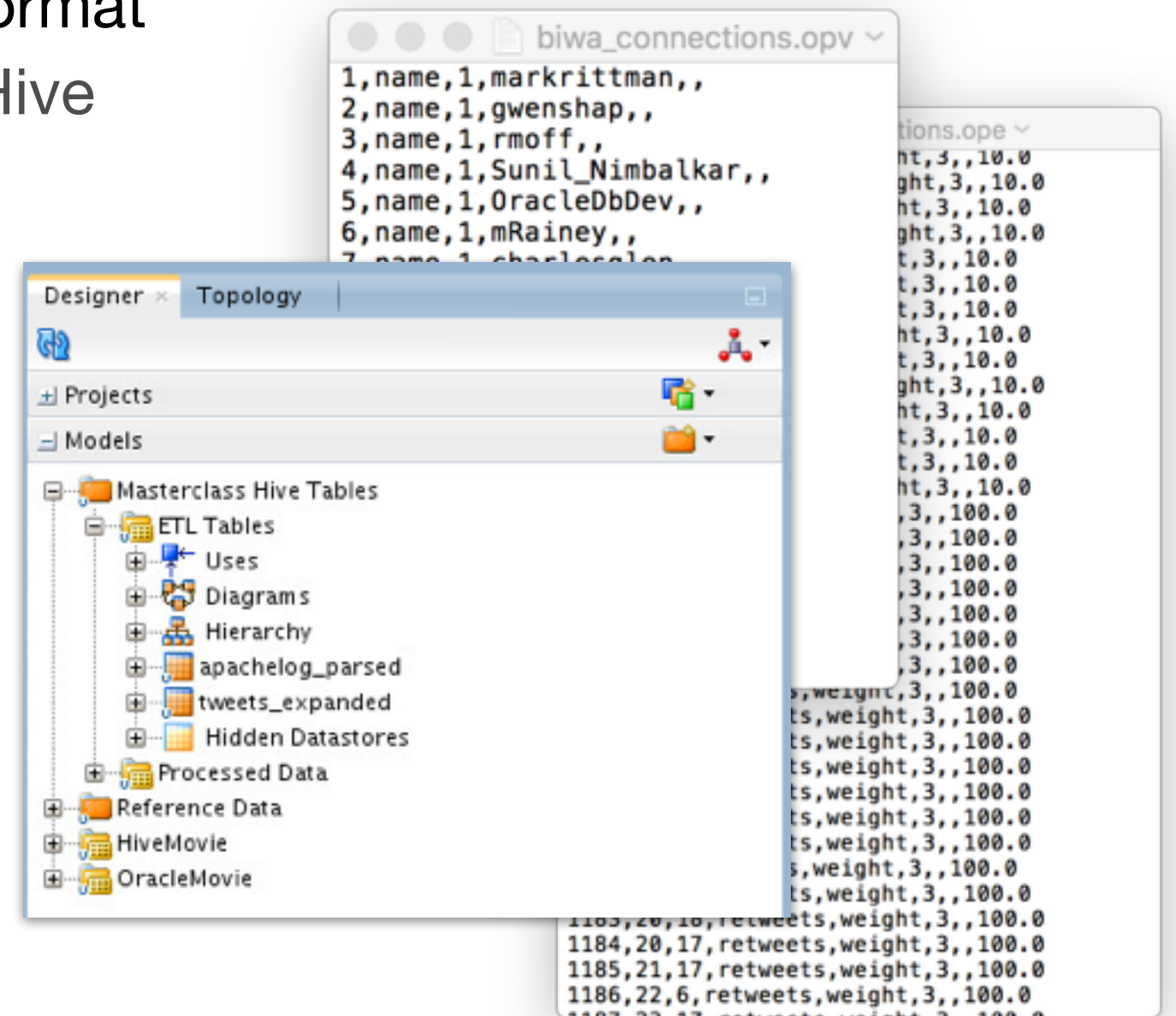
Massively Scalable Graph Store

- Oracle NoSQL
- HBase



Preparing Vertices and Edges for Ingestion

- 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)

```
biwa_connections.opv
1,name,1,markrittman,,
2,name,1,gwenshap,,
3,name,1,rmoff,,
4,name,1,Sunil_Nimbalkar,,
5,name,1,OracleDbDev,,
6,name,1,mRainey,,
7,name,1,charlesglen,,
8,name,1,Data88Geek,,
9,name,1,EricdeMarylebon,,
10,name,1,jpiwovar,,
11,name,1,retweetjava,,
12,name,1,Venkatram_T,,
13,name,1,NoSQLDigest,,
14,name,1,Data88Smart,,
15,name,1,i_m_dave,,
16,name,1,BizADave,,
17,name,1,Nephentur,,
18,name,1,HeliFromFinland,,
19,name,1,pulsebase,,
20,name,1,OsamaOracle,,
```

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

Edge File (.ope)

```
biwa_connections.ope
1153,3,101,mentions,weight,3,,10.0
1154,52,101,mentions,weight,3,,10.0
1155,6,101,mentions,weight,3,,10.0
1156,37,101,mentions,weight,3,,10.0
1157,6,37,mentions,weight,3,,10.0
1158,1,37,mentions,weight,3,,10.0
1159,17,3,mentions,weight,3,,10.0
1160,87,37,mentions,weight,3,,10.0
1161,2,93,mentions,weight,3,,10.0
1162,65,101,mentions,weight,3,,10.0
1163,3,101,mentions,weight,3,,10.0
1164,3,10,mentions,weight,3,,10.0
1165,37,6,mentions,weight,3,,10.0
1166,37,93,mentions,weight,3,,10.0
1167,2,1,retweets,weight,3,,100.0
1168,3,1,retweets,weight,3,,100.0
1169,4,1,retweets,weight,3,,100.0
1170,5,1,retweets,weight,3,,100.0
1171,7,6,retweets,weight,3,,100.0
1172,8,7,retweets,weight,3,,100.0
1173,9,7,retweets,weight,3,,100.0
1174,3,7,retweets,weight,3,,100.0
1175,10,6,retweets,weight,3,,100.0
1176,11,7,retweets,weight,3,,100.0
1177,12,7,retweets,weight,3,,100.0
1178,13,7,retweets,weight,3,,100.0
1179,14,7,retweets,weight,3,,100.0
1180,1,16,retweets,weight,3,,100.0
1181,3,6,retweets,weight,3,,100.0
1182,7,17,retweets,weight,3,,100.0
1183,20,18,retweets,weight,3,,100.0
1184,20,17,retweets,weight,3,,100.0
1185,21,17,retweets,weight,3,,100.0
1186,22,6,retweets,weight,3,,100.0
```

- 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

```
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);
```

Top 10 vertices

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

```
==>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

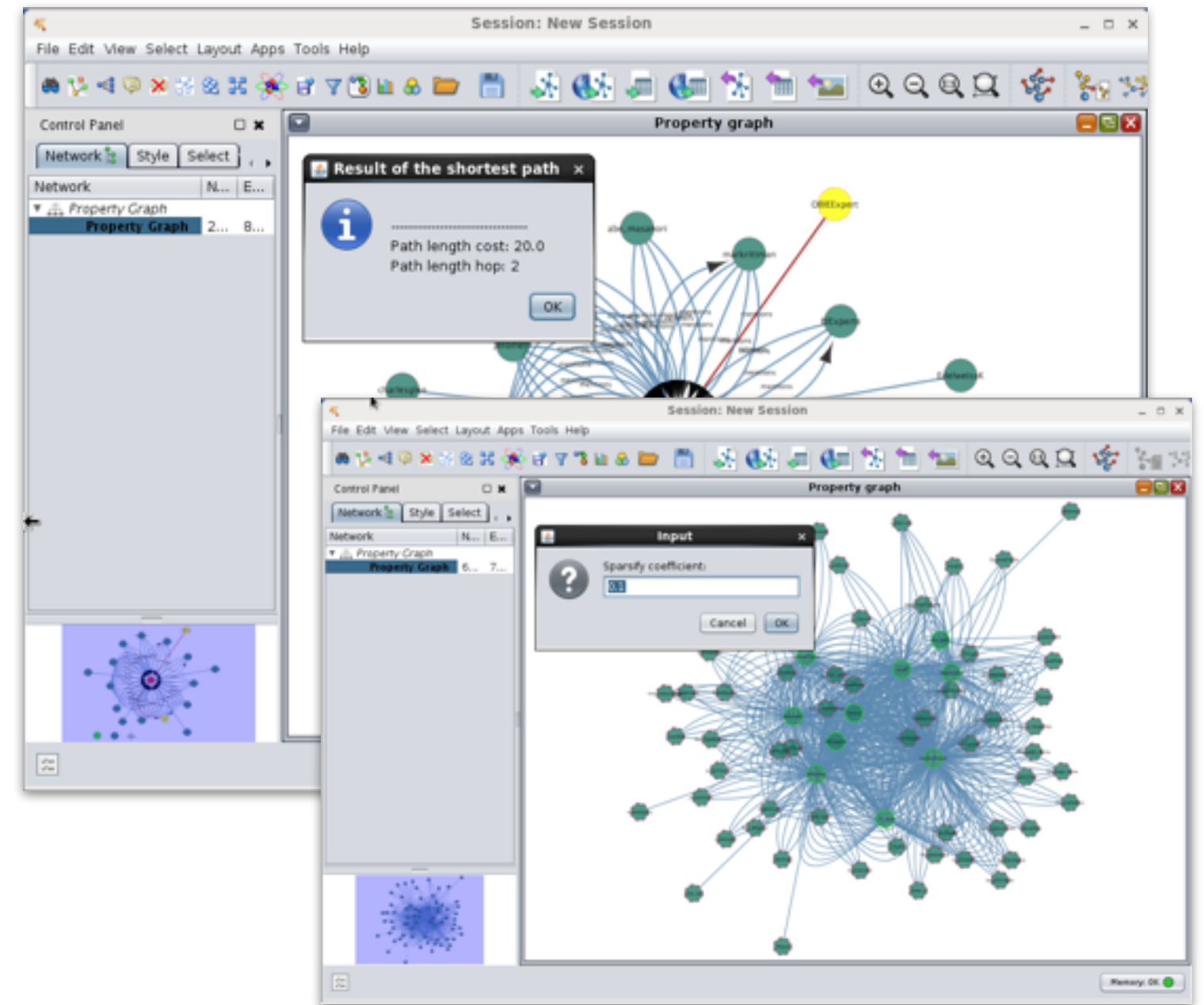
```
v1=opg.getVertex(11); v2=opg.getVertex(31); v3=opg.getVertex(1011); \
v4=opg.getVertex(61); v5=opg.getVertex(371); v6=opg.getVertex(171); \
v7=opg.getVertex(291); v8=opg.getVertex(651); v9=opg.getVertex(151); \
v10=opg.getVertex(931);
System.out.println("Top 10 influencers: \n " + v1.getProperty("name")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

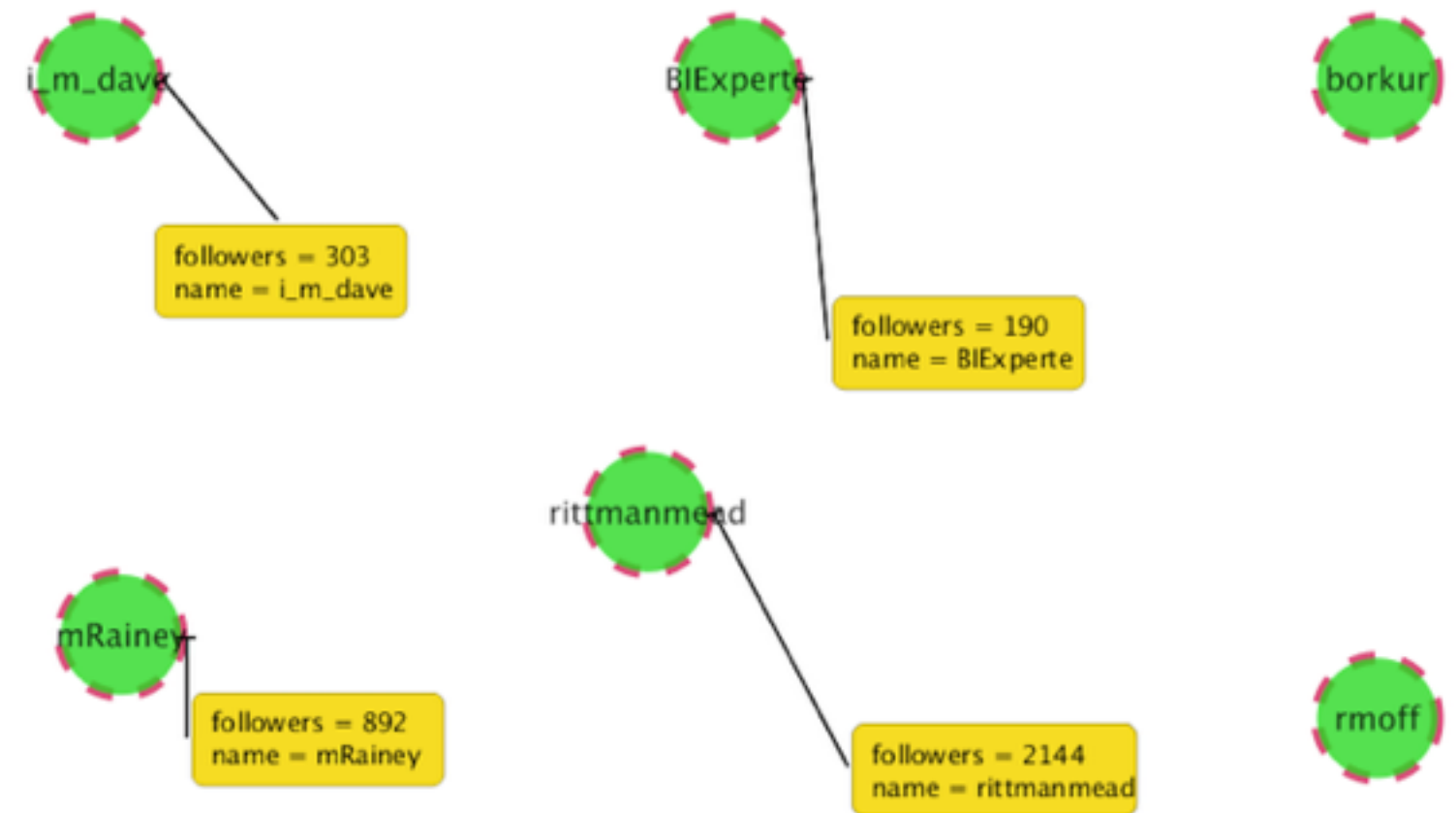
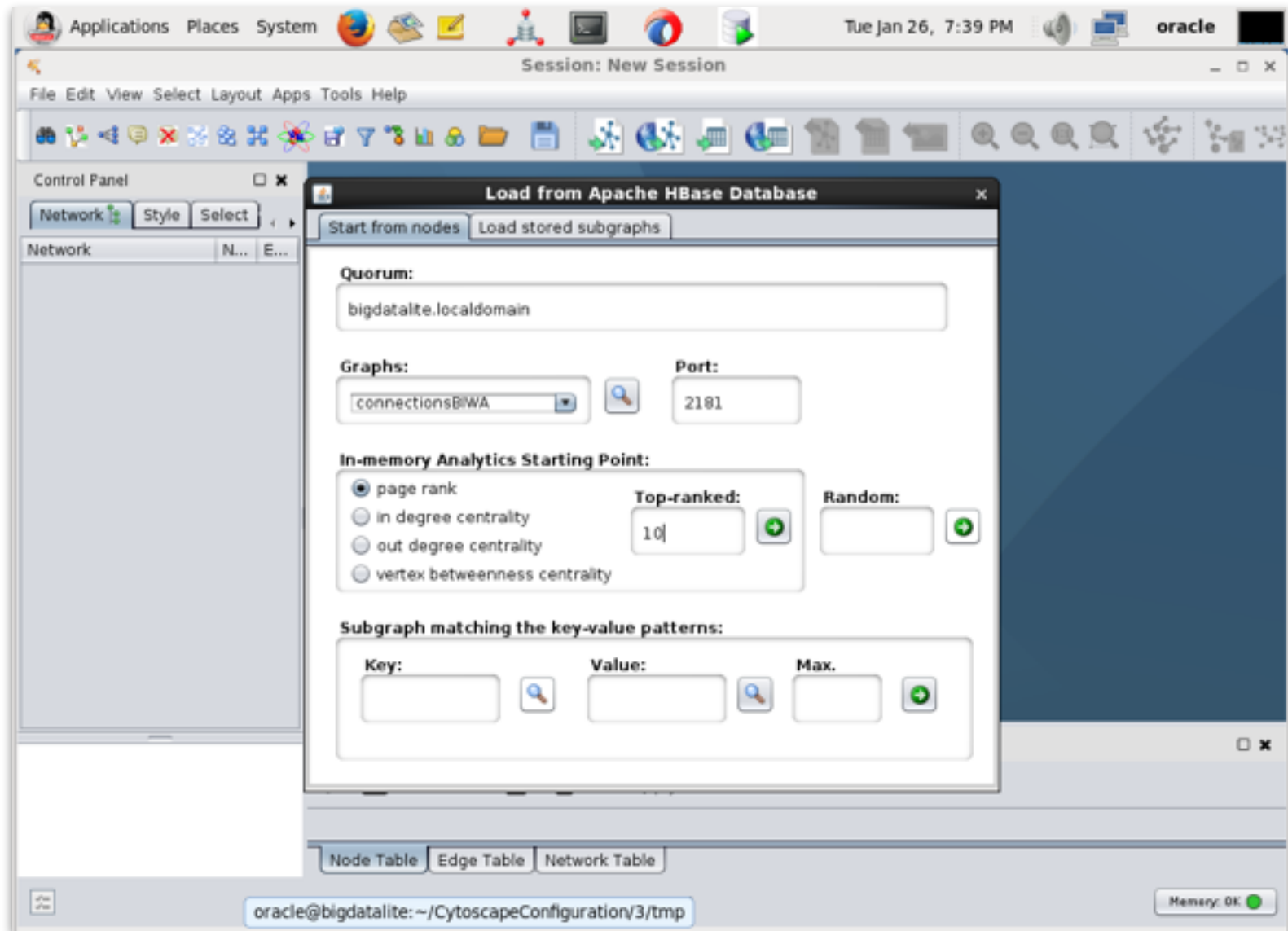
```
Top 10 influencers:
markrittman
rmoff
rittmanmead
mRainey
JeromeFr
Nephentur
borkur
BIExperte
i_m_dave
dw_pete
```

Visualising Property Graphs with Cityscape

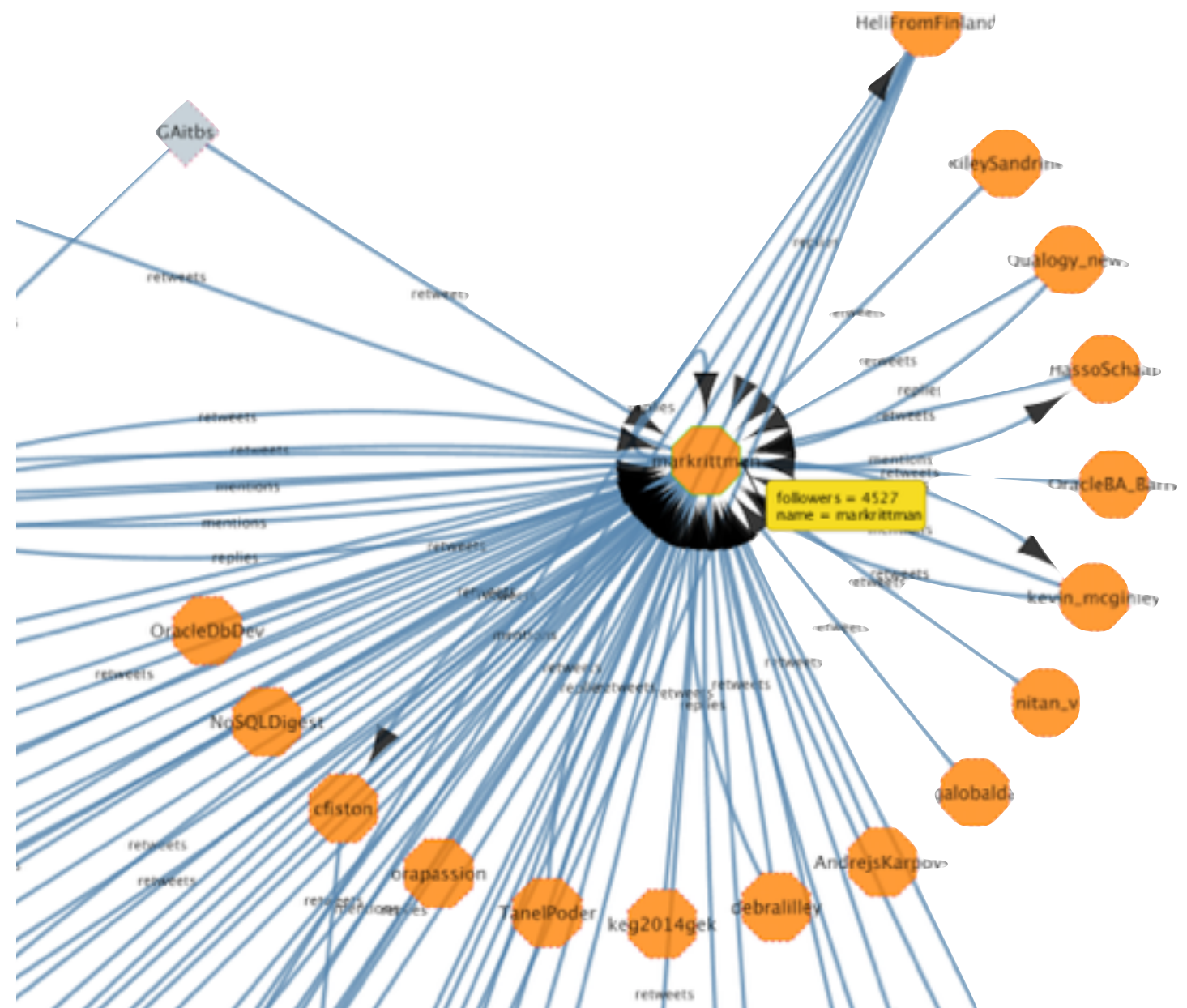
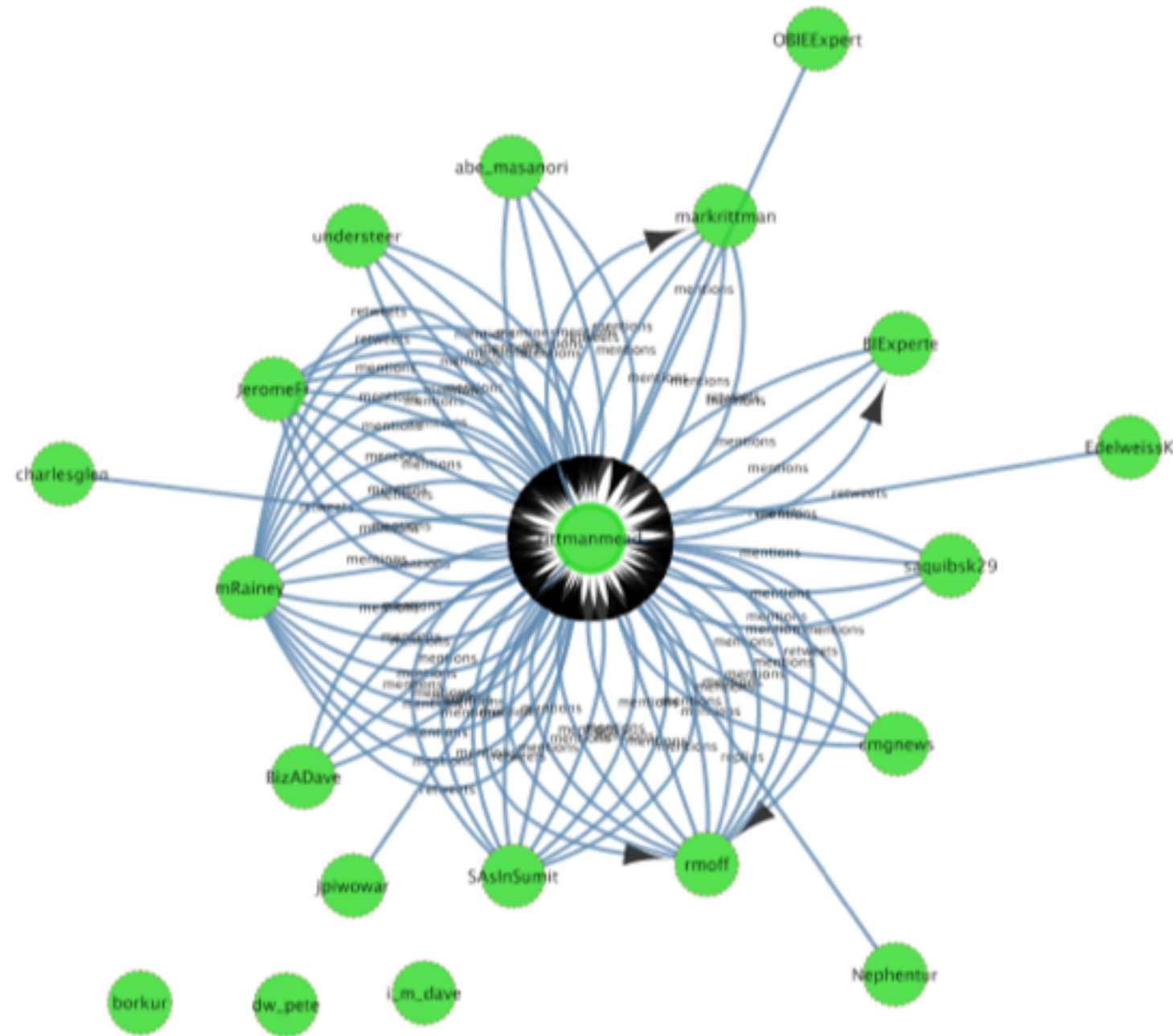
- Open source graph analysis tool with Oracle Big Data Graph and Spatial Plug-in
- Available shortly from Oracle, connects to Oracle NoSQL or HBase and runs Page Rank etc
- Alternative to command-line for In-Memory Analytics once base graph created



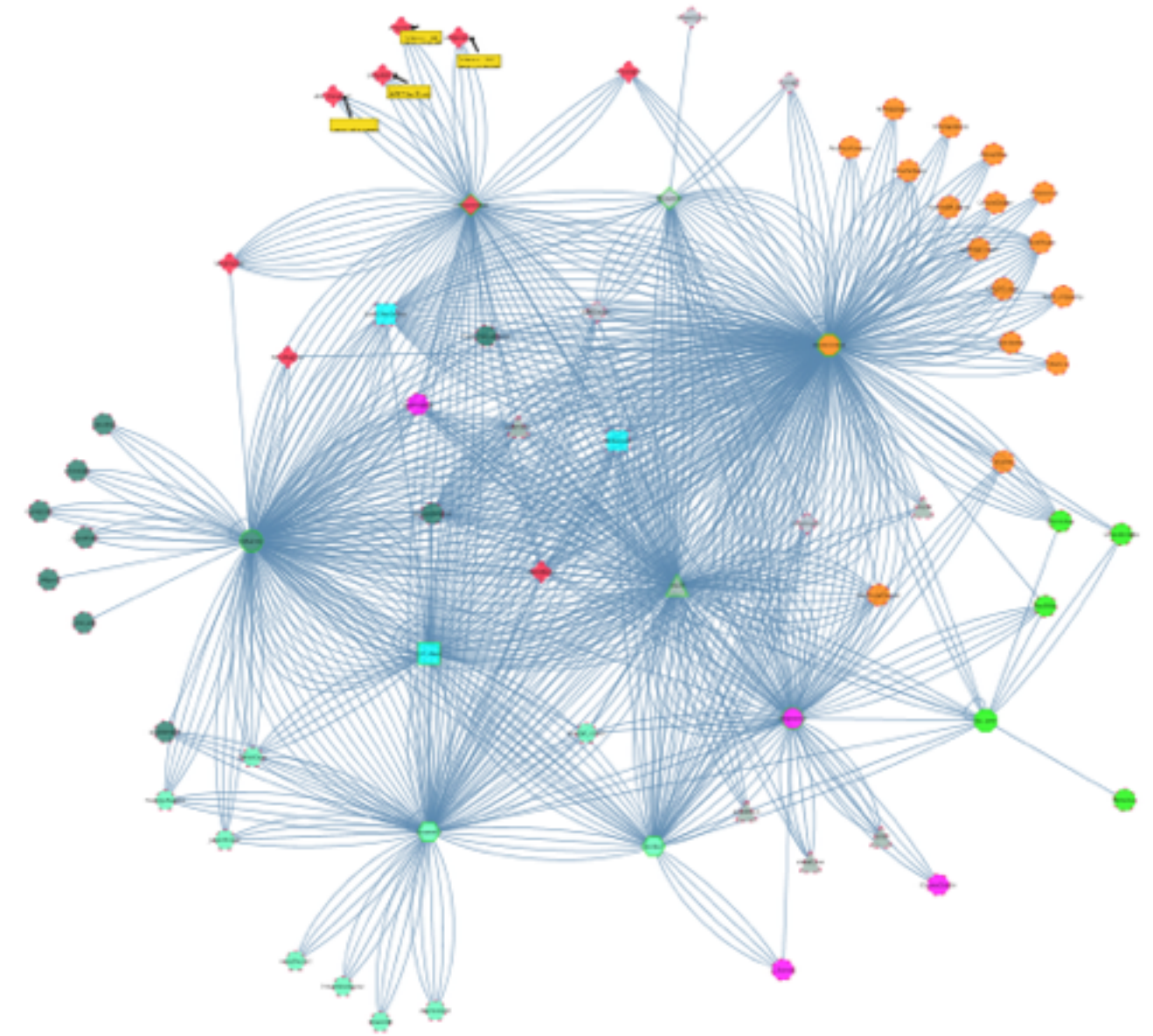
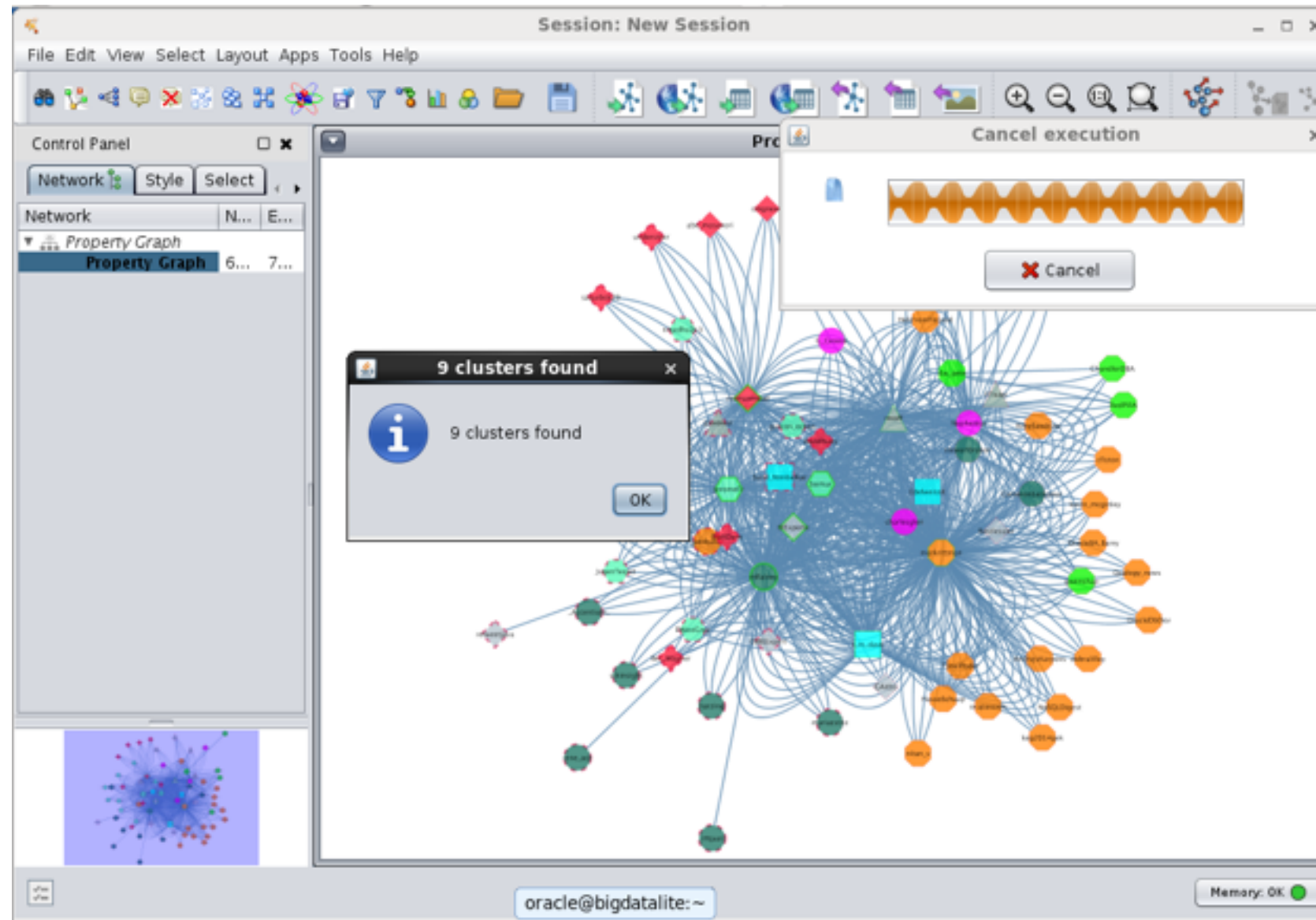
Calculating Top 10 Users using Page Rank Algorithm



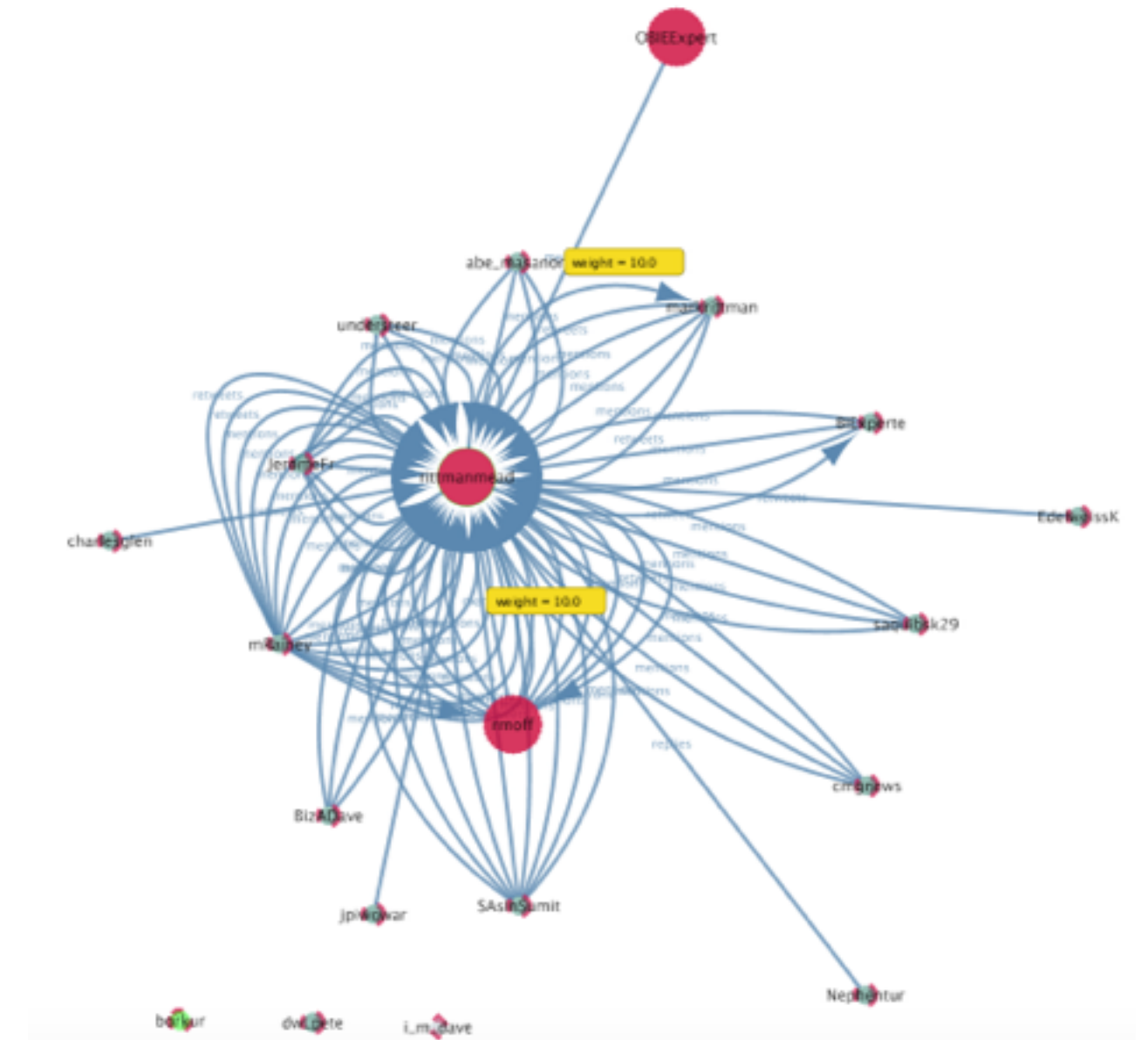
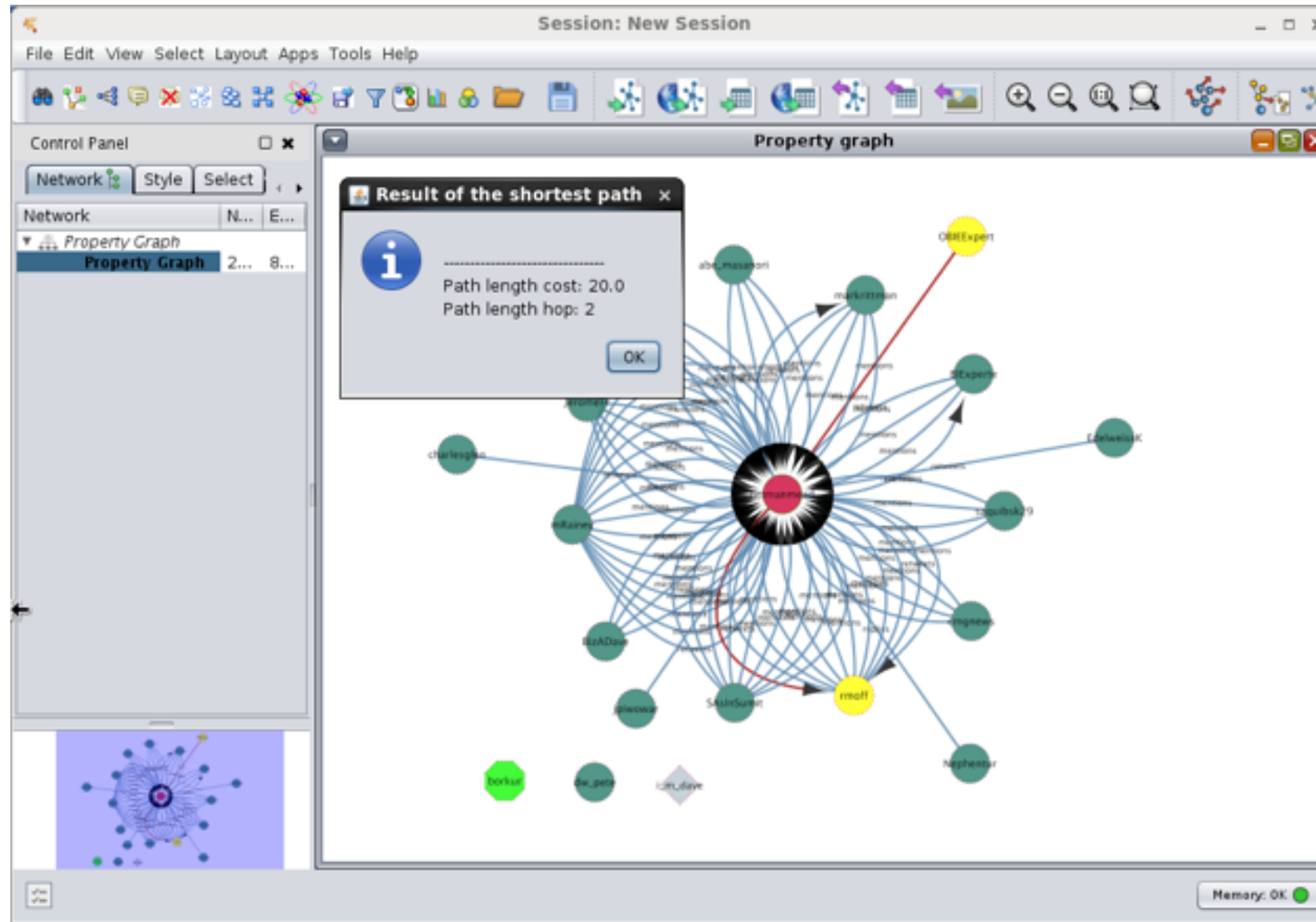
Visualising the Social Graph Around Particular Users



Detecting Clusters (Communities)

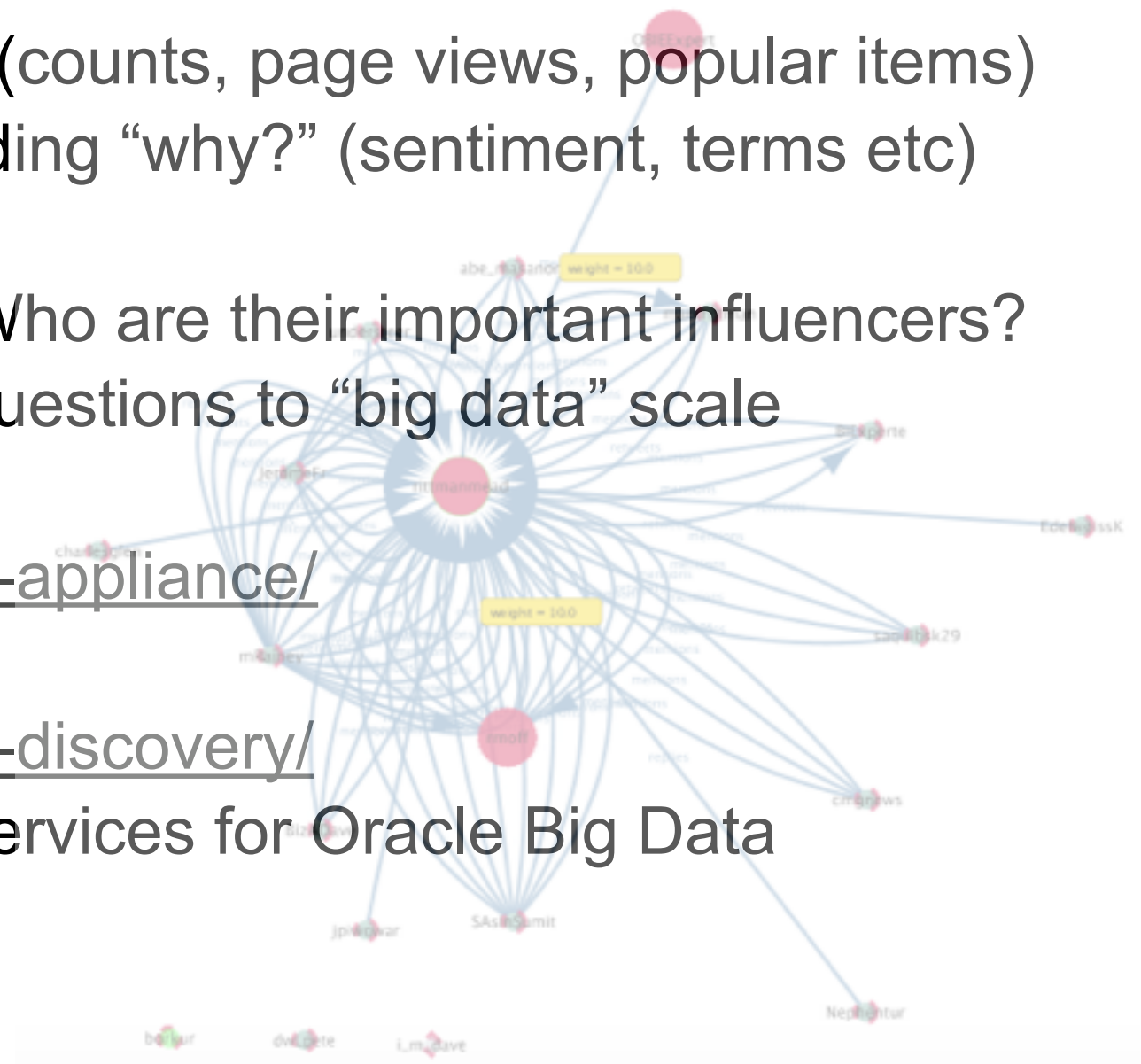


Calculating Shortest Path Between Users



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-appliance/>
 - ▶ <http://www.rittmanmead.com/category/big-data/>
 - ▶ <http://www.rittmanmead.com/category/oracle-big-data-discovery/>
- Rittman Mead offer consulting, training and managed services for Oracle Big Data
 - ▶ <http://www.rittmanmead.com/bigdata>



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

Mark Rittman, CTO, Rittman Mead
OTN EMEA Tour, May 2016