BIG Data Meetup

12.09.2012 – EMC, Beer-Sheva

Oshry Ben-Harush
Data Scientist, EMC IT
BIG Data

• How big is BIG data?

2009: 0.8 Zettabytes
Growing 44 Times

2020: 35.2 Zettabytes

Zettabyte \(2^{70}\) = 1024 Exabyte = 1024 Petabyte = 1024 Terabyte = 1024 Gigabyte = 1024 Megabyte
BIG Data – Challenges and Opportunities

• Two main challenges are tied with BIG data:
  – Storage – backup, de-duplication, disaster recovery, ...
  – Analysis – memory, compute

• Opportunities:
  – Gain illusive insights
  – Insights over time
  – Statistics
  – **Business value**
  – **Scientific value**
BIG Data – Illustrative Example – Twitter Topic Based Search

• Suppose that we are asked by Twitter to provide them with a text classification model to be incorporated in their search engine

• We are given 100,000 labeled twitter feeds classified to 20 topics for training the text classification model

• We are asked to:
  – Classify all of the currently stored tweets and to classify all future tweets on a daily basis
BIG Data – Illustrative Example – Twitter Topic Based Search

• Some details about twitter (Wikipedia):

“Twitter was created in March 2006 by Jack Dorsey and launched that July. The service rapidly gained worldwide popularity, with over **500 million active users as of 2012**, generating over **340 million tweets daily** and handling over **1.6 billion search queries per day**. Since its launch, Twitter has become one of the top 10 most visited websites on the Internet, and has been described as "the SMS of the Internet." Unregistered users can read tweets, while registered users can post tweets through the website interface, SMS, or a range of apps for mobile devices.”
BIG Data – Illustrative Example – Twitter Topic Based Search

• Outline:

  - Labeled tweets
  - Unlabeled Tweets
  - Model
  - Classification Model
  - Classify
  - Feedback
BIG Data – Illustrative Example – Twitter Topic Based Search

• Building the topic predictor is a bit more evolved. Model fitting:
BIG Data – Illustrative Example – Twitter Topic Based Search

• Model application:
BIG Data – Illustrative Example – Twitter Topic Based Search

- Twitter topic based search illustrates the difficulties that arise when trying to analyze BIG data:
  - Large volume of storage requirement
  - Highly capable computation device
  - Scale up requirement
  - Independent and repetitive tasks
BIG Data – Processing BIG data

- Move data to the compute for processing:

- Bring the compute to the data:
BIG Data – Processing BIG data

• Compute cluster using shared storage:

  ![Diagram of compute cluster using shared storage]

• Moving the compute into a distributed data storage and let each node perform the computation over its own data:

  ![Diagram of distributed data storage]

© Copyright 2012 EMC Corporation. All rights reserved.
BIG Data – Processing BIG data

• EMC has two enterprise scale solutions that take our discussion into account are Greenplum Hadoop and Greenplum database.

• Both are distributed, scalable, MPP systems that use work load sharing to store and process large amounts of data.
BIG Data Solutions – Greenplum Database

- Greenplum database – built on PostgreSQL
- Greenplum introduces the following improvements to PostgreSQL:
  - Massively Multi Parallel
  - Shared nothing infrastructure
  - Scalability
  - Within database deep analytics
BIG Data Solutions – Greenplum Database

• Schematic architecture:
BIG Data Solutions – Greenplum Database

• MAD skills for Enterprise Data warehouses states that it should be:
  – Magnetic – Enable the incorporation of new data as soon as it becomes available
  – Agile – Continues and rapid evolution of the database’s physical and logical content
  – Deep – Modern data warehouse should serve as both data repository and a sophisticated algorithmic runtime engine

MADlib
BIG Data Analytics – Examples

• Two examples will be used to illustrate BIG data analytics:
  – K-Means clustering using synthetic data – demonstrate Greenplum’s parallel processing capabilities within a data analytics setting
  – Text classification example – illustrates full text classification capabilities, all done within database
BIG Data Analytics – Examples – K-Means Clustering

• K-Means clustering
  – Assume an apple packaging factory
  – Three kinds of apples – measured for normalized color and radius 😊
  – The factory has a machine that automatically sorts the apples as these randomly enter
  – It uses a pre-defined threshold
  – After a while, 30,000,000 apples were packaged and the sorting machine needs to be re-calibrated

10,000 points randomly selected and imported into R for visualization and processing
BIG Data Analytics – Examples – K-Means Clustering

• The inputs to the K-Means algorithm:
  \[ D_{N \times 2} \] Data matrix for N samples in two dimensions
  \[ K \] The number of center points we would like to find

• K-Means data stored in the database has the following structure

<table>
<thead>
<tr>
<th>Dist_key</th>
<th>Id</th>
<th>Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integer</td>
<td>Integer</td>
<td>Float8[]</td>
</tr>
<tr>
<td>The table distribution key</td>
<td>sample identifier</td>
<td>two dimensional arrays of data</td>
</tr>
</tbody>
</table>
BIG Data Analytics – Examples – K-Means Clustering

1. Randomly select K points from the data set these as the initial centers $C^1_{K \times 2}$

Initial centroids selected from the 10,000 randomly selected samples (in R)
BIG Data Analytics – Examples – K-Means Clustering

2. For each data point find the closest center $P_{Nx1}$ (by some measure)

3. Calculate new centers $C_{Kx2}^{i+1}$ by averaging the points assigned to each center $C_{Kx2}^i$

Centroids following a single K-Means (in R) iteration over the 10,000 random samples
BIG Data Analytics – Examples – K-Means Clustering

4. Repeat steps 2 and 3 for the required number of iterations or until the centers do not move
BIG Data Analytics – Examples – K-Means Clustering

- K-Means clustering is performed using a single call to a MADlib function

```sql
SELECT * FROM madlib.kmeans_plusplus('kmeans_data', 'data', 'id', 'kmeans_points', 'kmeans_centroids', 'l2norm', 4, 0.001, FALSE, FALSE, 3, 0.01);
```
BIG Data Analytics – Examples – K-Means Clustering

- The K-Means algorithm is attempted twice:
  - Using a parallel processing – the data table is distributed across all segment servers
  - Using a single segment server – the data table resides in a single segment server – high data skew

- Timing (4 iterations):
  - Parallel processing: **46.03 Sec**
  - Single segment server processing: **169.43 Sec**
BIG Data Analytics – Examples – K-Means Clustering

• K-Means clustering results:

Why BIG data?
Why not always sample?
BIG Data Analytics – Examples – K-Means Clustering

Curse of dimensionality

DATA MATTERS

Yoshua Bengio
Full Professor
Department of Computer Science and Operations Research
Canada Research Chair in Statistical Learning Algorithms
<my first name> (dot) <my last name> (at)umontreal (dot) ca
BIG Data Analytics – Examples – Text classification

• For text classification, the 20 Newsgroups data set was chosen: http://qwone.com/~jason/20Newsgroups/

• This dataset contains 18,846 posts collected from news groups discussing 20 topics:

<table>
<thead>
<tr>
<th>comp.graphics</th>
<th>rec.autos</th>
<th>talk.politics.misc</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp.os.ms-windows.misc</td>
<td>rec.motorcycles</td>
<td>talk.politics.guns</td>
</tr>
<tr>
<td>comp.sys.ibm.pc.hardware</td>
<td>rec.sport.baseball</td>
<td>talk.politics.mideast</td>
</tr>
<tr>
<td>comp.sys.mac.hardware</td>
<td>rec.sport.hockey</td>
<td></td>
</tr>
<tr>
<td>comp.windows.x</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>misc.forsale</td>
<td>talk.religion.misc</td>
</tr>
<tr>
<td></td>
<td>sci.crypt</td>
<td>alt.atheism</td>
</tr>
<tr>
<td></td>
<td>sci.electronics</td>
<td>soc.religion.christian</td>
</tr>
<tr>
<td></td>
<td>sci.med</td>
<td></td>
</tr>
<tr>
<td></td>
<td>sci.space</td>
<td></td>
</tr>
</tbody>
</table>

20 Newsgroups by topic

• The data set is split into samples that are meant for model training and to samples that are meant for testing
BIG Data Analytics – Examples – Text classification

• Text classification setting:

- Labeled docs
- Train set
- Test Set
- Train set features
- Test features
- Dictionary
- Preprocessing and feature extraction
- Model
- Predict and score
BIG Data Analytics – Examples – Text classification

CALL FOR PRESENTATIONS

NAVY SCIENTIFIC VISUALIZATION AND VIRTUAL REALITY SEMINAR

Tuesday, June 22, 1993
Carderock Division,
Naval Surface Warfare Center
(formerly the
David Taylor Research Center)
Bethesda, Maryland

SPONSOR: NESS (Navy Engineering Software System) is sponsoring a one-day Navy Scientific Visualization ...

ABOUT THIS POSTING
---------------------

This is a (still rather rough) listing of likely questions and information about RIPEM, a program for public key mail encryption. It (this FAQ, not RIPEM) was written and will be maintained by Marc VanHeyningen, <mvanheyen@whale.cs.indiana.edu>. It will be posted to a variety of newsgroups on a monthly basis; follow-up discussion specific to RIPEM is redirected to the group alt.security.ripem.

This month, I have reformatted this posting in ...

Atheist Resources

Addresses of Atheist Organizations

USA

FREEDOM FROM RELIGION FOUNDATION

Darwin fish bumper stickers and assorted other atheist paraphernalia are available from the Freedom From Religion Foundation in the US.

Write to: FFRF, P.O. Box 750, Madison, WI 53701.
Tel ...

Comp.Graphics

Sci.Crypt

Alt.Atheism
BIG Data Analytics – Examples – Text classification

Pre-processing and tokenization

Comp. Graphics

{call, for, presentations, navy, scientific, visualization, and, virtual, reality, seminar, tuesday, june, num, num, carderock, division, naval, surface, warfare, center, formerly, the, david, taylor, research, center, bethesda, maryland, sponsor, ness, navy, emerging, software, system, sponsoring, one, day, navy, scientific, visualization, and, virtual, reality, seminar, the, purpose, the, seminar, present, and, exchange, information, for, navy, related, scientific, visualization, and, virtual, reality, programs, research, developments, and, applications}

Sci. Crypt

{about, this, posting, this, still, rather, rough, listing, likely, questions, and, information, about, ripem, program, for, public, key, mail, encryption, this, faq, not, ripem, was, written, and, will, maintained, marc, vanheynigen, mvanheyn, whale, indiana, edu, will, posted, variety, newsgroups, monthly, basis, follow, discussion, specific, ripem, redirected, the, group, alt, security, ripem, this, month, have, reformatted, this, posting, attempt, comply, with, the, standards, for, hypertext,faq, formatting, allow, easy, manipulation, this, document, over, the, world, wide, web, let, know, what, you, think}

Alt. Atheism

{atheist, resources, addresses, atheist, organizations, usa, freedom, from, religion, foundation, darwin, fish, bumper, stickers, and, assorted, other, atheist, paraphernalia, are, available, from, the, freedom, from, religion, foundation, the, write, ffrf, box, num, madison, num, telephone, num, num, num, evolution, designs}
BIG Data Analytics – Examples – Text classification

Create a dictionary of unique words from all of the documents used for training

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>taylor, research, center, bethesda, maryland, sponsor, ness, navy, engineering, software, system, sponsoring, one, day, navy, scientific, visualization, seminar, the, en, mvanheyn, whale, indiana, edu, will, posted, variety, newsgroups, monthly, basis, follow, discussion, specific, ripem, redirected, the, group, alt, security, ripem, this, month, have, om, religion, foundation, darwin, fish, bumper, stickers, and, assorted, other, atheist, paraphernalia, are, available, from, the, freedom, from, religion, foundation, the, write, ffrf, box,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

{about, addresses, allow, alt, and, applications, are, assorted, atheist, attempt, available, basis, bethesda, box, bumper, call, car derock, center, comply, darwin, david, day, designs, developments, discussion, division, document, easy, edu, encryption, engineering, evolution, exchange, faq, ffrf, fish, follow, for, formatting, formerly, foundation, freedom, from, group, have, hypertext, in diana, information, june, key, know, let, likely, listing, madison, mail, maintained, manipulation, marc, maryland, month, monthl y, mvanheyn, naval, navy, ness, newsgroups, not, num, one, organizations, other, over, paraphernalia, posted, posting, present, presentations, program, programs, public, purpose, questions, rather, reality, redirected, reformatted, related, religion, research, resources, ripem, rough, scientific, security, seminar, software, specific, sponsor, sponsoring, standards, stickers, still, surface, e, system, taylor, telephone, the, think, this, tuesday, usa, vanheyningen, variety, virtual, visualization, warfare, was, web, whale , what, wide, will, with, world, write, written, you}
Map the documents using the dictionary to the feature space

\[
\{\text{address, allow, alt, and, applications, are, assorted, atheist, attempt, available, basis, bethesda, box, bumper, call, car, derock, center, comply, darwin, day, designs, developments, discussion, division, document, easy, edu, encryption, engineering, evolution, exchange, faq, ffrf, fish, follow, for, formatting, formerly, foundation, freedom, from, group, have, hypertext, in, diana, information, june, key, know, let, likely, listing, madison, mail, maintained, manipulation, marc, maryland, month, monthly, mvanheyn, naval, navy, ness, newsgroups, not, num, one, organizations, other, over, paraphernalia, posted, posting, present, presentations, program, programs, public, purpose, questions, rather, reality, redirected, reformatted, related, religion, research, resources, ripem, rough, scientific, security, seminar, software, specific, sponsor, sponsoring, standards, stickers, still, surface, system, taylor, telephone, the, think, this, tuesday, usa, vanheyningen, variety, virtual, visualization, warfare, was, web, whale, what, wide, will, with, world, write, written, you}\}
\]
BIG Data Analytics – Examples – Text classification

Train a classification model for each class using the features and labels

Comp.Graphics Model

Sci.Crypt Model

Alt.Atheism Model

{0,0,0,0,5,1,0,0,0,0,0,0,1,0,0,1, 1,2,0,0,1,1,0,1,0,0,0,0,0,1,0,1, 0,0,0,0,0,2,0,1,0,0,0,0,0,0,0,1,1,0, 0,0,0,0,0,0,0,0,0,1,0,0,1,4,1,0, 0,2,1,0,0,0,0,0,1,1,0,1,0,1,0,0, 3,0,0,1,0,2,0,0,0,3,0,3,1,0,1,1,0, 0,0,1,1,1,0,3,0,0,1,0,0,0,3,3,1,0, 0,0,0,0,0,0,0,0,0,0}  

{2,0,1,1,2,0,0,0,0,1,0,1,0,0,0,0,0, 0,0,1,0,0,0,0,0,1,0,1,1,1,0,0,0, 2,0,0,1,2,1,0,0,0,0,1,1,1,1,1,0,1, 1,1,1,1,0,1,1,1,1,0,1,1,1,0,0,1, 1,0,0,0,0,1,0,1,2,0,0,1,0,1,0,1, 1,0,1,0,0,0,0,4,1,0,1,0,0,1,0,1, 0,1,0,0,0,0,3,1,6,0,0,1,1,0,0,0,1, 1,1,1,1,2,1,1,0,1,1}  

{0,1,0,0,1,0,1,1,3,0,1,0,0,1,0,1,0, 0,0,0,1,0,0,1,0,0,0,0,0,0,0,0,0,1,0, 0,1,1,0,0,0,2,2,3,0,0,0,0,0,0,0, 0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0, 0,5,0,1,1,0,1,0,0,0,0,0,0,0,0,0, 0,0,0,0,2,0,1,0,0,0,0,0,0,0,0,0, 1,0,0,0,0,1,2,0,0,0,1,0,0,0,0,0, 0,0,0,0,0,0,0,0,1,0,0}
BIG Data Analytics – Examples – Text classification

- One final step – One-VS-All classification
BIG Data Analytics – Examples – Text classification

- The data is loaded into the Greenplum server using ODBC through python
- The data is loaded into two tables, one is used for training the SVM model and the other is used for testing. Both tables have the same structure:

<table>
<thead>
<tr>
<th></th>
<th>Id (bigint)</th>
<th>Class (text)</th>
<th>Content (text)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A unique identifier for each document</td>
<td>Class label – extracted from the directory structure</td>
<td>Textual content</td>
</tr>
</tbody>
</table>
BIG Data Analytics – Examples – Text classification

• Once the data is loaded the following processing is performed:
  1. Create a view that contains classes mapping to integers for both train and test subsets

  -- *Create a view that contains the classes id*
  CREATE OR REPLACE VIEW tng_classes_view AS
  SELECT row_number() OVER() as class_id, * FROM
  (SELECT DISTINCT class FROM tng_data_train) AS classes;

  -- *Create a view that also contains the class ID*
  CREATE OR REPLACE VIEW tng_data_train_view AS
  SELECT tng_data_train.id, tng_data_train.class,
  tng_classes_view.class_id AS class_id, tng_data_train.content
  FROM tng_data_train, tng_classes_view
  WHERE tng_data_train.class = tng_classes_view.class;

  CREATE OR REPLACE VIEW tng_data_test_view AS
  SELECT tng_data_test.id, tng_data_test.class,
  tng_classes_view.class_id AS class_id, tng_data_test.content
  FROM tng_data_test, tng_classes_view
  WHERE tng_data_test.class = tng_classes_view.class;
BIG Data Analytics – Examples – Text classification

2. Pre-process the text data:

```sql
-- Create a preprocessed view of the data - strings are stripped from
-- unrequired characters and converted to arrays
CREATE OR REPLACE VIEW tng_data_train_array_view
AS SELECT id, class, class_id,
string_to_array(
regexp_replace(
regexp_replace(
regexp_replace(
regexp_replace(
regexp_replace(
regexp_replace(
      lower(content),
      E'\W', ' ', 'g'),
      E'\d+', ' num ', 'g'),
      E'(\y\w{1,2}\y|\y\w{30,}\y)', ' ', 'g'),
      E'\s+', ' ', 'g'),
      ' $', ''),
      E'^\s+', ' ', 'g'),
      ' ')
) as content from tng_data_train_view;
```
BIG Data Analytics – Examples – Text classification

3. Create the dictionary from training set

    -- Create a dictionary view of the data
    CREATE OR REPLACE VIEW tng_dictionary_view
    AS SELECT array_agg(words ORDER BY words) as words FROM (SELECT DISTINCT unnest(content) as words
    FROM tng_data_train_array_view ORDER BY WORDS) foo;

4. Use MADlib to extract sparse feature vectors from the text content given the dictionary

    CREATE TABLE tng_sfv_train_data AS
    SELECT id, class, class_id as label, madlib.svec_sfv((select * FROM tng_dictionary_view limit 1), content)::float8[]
    AS ind from tng_data_train_array_view DISTRIBUTED BY(id);

    CREATE TABLE tng_sfv_test_data AS
    SELECT id, class, class_id as label, madlib.svec_sfv((select * FROM tng_dictionary_view limit 1), content)::float8[]
    AS ind from tng_data_test_array_view DISTRIBUTED BY(id);
BIG Data Analytics – Examples – Text classification

5. Create views for training the classifier of class 1 (sci.crypt) from the rest

```
CREATE OR REPLACE VIEW tng_sfv_train_sci_crypt_view AS
SELECT id, ind,
       CASE WHEN label = 1 THEN 1 ELSE -1 END AS label
FROM tng_sfv_train_view;

CREATE OR REPLACE VIEW tng_sfv_test_sci_crypt_view AS
SELECT id, ind,
       CASE WHEN label = 1 THEN 1 ELSE -1 END AS label
FROM tng_sfv_test_view;
```

6. Use MADlib to train the SVM model

```
SELECT madlib.lsvm_classification('text_demo.tng_sfv_train_sci_crypt_view',
                                  'text_demo.tng_svm_sci_crypt', false,
                                  true, 0.1, 0.001);
```
BIG Data Analytics – Examples – Text classification

7. Create classification tables for the train and test set

```sql
SELECT madlib.lsvm_predict_batch('text_demo.tng_sfv_train_sci_crypt_view', 'ind', 'id', 'text_demo.tng_svm_sci_crypt', 'text_demo.tng_svm_sci_crypt_predict_train', false);

SELECT madlib.lsvm_predict_batch('text_demo.tng_sfv_test_sci_crypt_view', 'ind', 'id', 'text_demo.tng_svm_sci_crypt', 'text_demo.tng_svm_sci_crypt_predict_test', false);
```

8. Score the model using the train and test data. For example, classification error for the train set:

```sql
SELECT sum((prediction_label!=label)::int)/count(*)::float8 AS train_classification_error FROM (SELECT tng_sfv_train_sci_crypt_view.id, label, prediction,
        CASE WHEN prediction > 0 THEN 1 ELSE -1 END AS prediction_label
    FROM tng_sfv_train_sci_crypt_view, tng_svm_sci_crypt_predict_train
    WHERE tng_sfv_train_sci_crypt_view.id = tng_svm_sci_crypt_predict_train.id) predict_view;
```
BIG Data Analytics – Examples – Text classification

• Results for this example will not provide good results and will be omitted, we have neglected:
  – Stop words elimination
  – Stemming
  – Feature selection
  – Dimensionality reduction
  – Cross validation – model parameter optimization
BIG Data Analytics – Conclusion

- K-Means: simple algorithm harnessing Greenplum’s parallel capabilities in an embarrassingly parallel setting

- Text: we have shown a complete text classification solution to be used within the database – for scientific or enterprise text mining – most of the tasks in this example were also embarrassingly parallel

- When attempting data analytics over BIG data, there is practically no alternative than to use a distributed, scalable computationally capable infrastructure

- The modern data warehouse is fully capable to both store/retrieve information and to apply deep analytics over it’s stored data
BIG Data Analytics – Conclusion

• Once computation capabilities are migrated into the data warehouse, the data analysts, data base administrators and data scientists can collaborate on the same framework using common data, but different approaches.
Thank You!