Sentiment Analysis using Hadoop

Sponsored By Atlink Communications Inc

Instructor : Dr. Sadegh Davari
Mentors : Dilhar De Silva, Rishita Khalathkar

Team Members : Ankur Uprit, Pinaki Ranjan Ghosh, Kiranmayi Ganti, Srijha Reddy Gangidi

Capstone Project Group 1
What is Sentiment Analysis?
Sentiment Analysis with Twitter Classification of Data
Types of Sentiment Analysis
Introduction to the Project
What is Hadoop and HDFS?
Structured and Unstructured Data

Ankur Uprit
Team Leader/ Application Developer

Capstone Project Group 1
Sentiment Analysis

- Sentiment analysis is the detection of **attitudes**
  - Enduring, affectively colored beliefs, dispositions towards objects or persons
  1. **Holder (source)** of attitude
  2. **Target (aspect)** of attitude
  3. **Type** of attitude
    - From a set of types
      - *Like, love, hate, value, desire,* etc.
    - Or (more commonly) simple weighted **polarity**:
      - *positive, negative, neutral,* together with **strength**
  4. **Text** containing the attitude
    - Sentence or entire document
Sentiment Analysis

- Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document.
- The attitude may be his or her:
  1. **Judgment**
  2. **Affective state** (that is to say, the emotional state of the author when writing)
  3. **Intended emotional communication** (that is to say, the emotional effect the author wishes to have on the reader)
Sentiment Analysis With Twitter

- twitter.com is a popular microblogging website
- Each tweet is 140 characters in length
- Tweets are frequently used to express a tweeter's emotion on a particular subject
- There are firms which poll twitter for analyzing sentiment on a particular topic
- The challenge is to gather all such relevant data, detect and summarize the overall sentiment on a topic
Classification Of Data

- Polarity classification – Positive
  - Negative Sentiment
- 3-way classification – Positive
  - Negative
  - Neutral
Types of sentiment analysis

- **Movie:** Is this review positive or negative?
- **Products:** What do people think about the new iPhone?
- **Public Sentiment:** How is consumer confidence? Is despair increasing?
- **Politics:** What do people think about this candidate or issue?
- **Prediction:** Predict election outcomes or market trends from sentiment
Introduction to the project

Sentiment Analysis Using Hadoop & Hive
**What is Hadoop and HDFS?**

**Hadoop:** A Software Framework for Data Intensive Computing Applications

- Software platform that lets one easily write and run applications that process vast amounts of data. It includes:
  - MapReduce – offline computing engine
  - HDFS – Hadoop distributed file system
  - HBase (pre-alpha) – online data access

- Yahoo! is the biggest contributor
What does Hadoop do?

- Hadoop implements Google’s MapReduce, using HDFS.
- MapReduce divides applications into many small blocks of work.
- HDFS creates multiple replicas of data blocks for reliability, placing them on compute nodes around the cluster.
- MapReduce can then process the data where it is located.
- Hadoop’s target is to run on clusters of the order of 10,000-nodes.
The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware.

It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant.

- **Highly fault-tolerant** and is designed to be deployed on low-cost hardware.
- **Provides high throughput access** to application data and is suitable for applications that have large data sets.
- Relaxes a few POSIX requirements to enable streaming access to file system data.
- Part of the Apache Hadoop Core project. The project URL is [http://hadoop.apache.org/core/](http://hadoop.apache.org/core/).
Sentiment Analysis Using Hadoop & Hive

- The twitter data is mostly unstructured
- Hadoop is the technology that is capable of dealing with such large unstructured data
- In this project, Hadoop Hive on Windows will be used to analyze data.
- This analysis will be shown with interactive visualizations using some powerful BI tools for Excel like Power View
- Finally, a real time case study will be used to create a report on how Sentiment Analysis can be implemented for a product
- What infrastructure, skills, technology would be most ideal and how it would help in improving the brand image/ quality of the product
Technologies Used

- HortonWorks Data Platform for Windows
- Hive and HiveQL
- BI tools for Excel

Research, Analysis and Design

- We had carried out a detail analysis on existing solutions in the market within the project scope
- Followed tutorials on YouTube
- Analyze the raw data, learned about unstructured data. How its been used and managed
Requirements Specification

Software Requirement Specification draft that includes a UML 2.0 use case, analysis and Sequence models.

Use Case Diagram

Sequence Diagram
Design Specification

- Software Design Specification includes a UML 2.0 design model and a data model

Test and Deliver

- Product Tests specified with final and working version of the application with unit testing and system testing.
What Is Structured Data?

- Data that resides in a fixed field within a record or file is called structured data including relational databases and spreadsheets.

- Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed.

- Structured data has the advantage of being easily entered, stored, queried and analyzed.

- At one time, because of the high cost and performance limitations of storage, memory and processing, relational databases and spreadsheets using structured data were the only way to effectively manage data.
What Is Unstructured Data?

Unstructured data, usually binary data that is proprietary, is that which has no identifiable internal structure.

Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, pdf files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

80% of business-relevant information originates in unstructured form, primarily text.
What is Hive?
Why Hive?
What is HiveQL?
HiveQL Operations?
What is Hortonworks Data Platform (HDP)?
HDP System Requirements
Setting HDP on Virtual Environment.

Pinaki Ranjan Ghosh
Application Developer / Designer

Capstone Project Group 1
Hive

- Tools to enable easy data extract/transform/load (ETL)
- A mechanism to impose structure on a variety of data formats
- Access to files stored either directly in HDFS or in other data storage systems
- Query execution via MapReduce

Large datasets stored in Hadoop's HDFS

Querying

Managing

Summarization

Analysis
Hive is a data-warehouseing infrastructure for Hadoop

Warehoused data

Easy to retrieve and Easy to manage.

The data are organized in three different formats in HIVE

- **Tables**: They are very similar to RDBMS tables and contains rows and tables.
- **Partitions**: Hive tables can have more than one partition like subdirectories and file systems.
- **Buckets**: Data may be divided into buckets which are stored as files in partition in the underlying file system.
HiveQL

- HiveQL is the Hive query language
- It is a SQL-like interface on top of Hadoop
- Hive converts queries written in HiveQL into MapReduce tasks that are then run across the Hadoop cluster to fetch the desired results

• Examples:
  1. Create TABLE sample_table (name String, age int);
  2. LOAD DATA LOCAL PATH ‘input/mydata/data.txt’ INTO TABLE mytable;
  3. Insert into birthday Select firstname, lastname, birthday from customers where birthday is NOT NULL;
  4. Select * from myTable;
HiveQL

• Create and manage tables and partitions
• Support various Relational, Arithmetic and Logical Operators
• Evaluate functions
• Download the contents of a table to a local directory or result of queries to HDFS directory

Main Operations...

ANALYZE TABLE
DESCRIBE COLUMN
DESCRIBE DATABASE
EXPORT TABLE
IMPORT TABLE
LOAD DATA
SHOW TABLE EXTENDED
SHOW INDEXES
SHOW COLUMNS
Hortonworks Data Platform (HDP)

• Hortonworks and Microsoft have partnered to bring the benefits of Apache Hadoop to Windows

• HDP provides an enterprise ready data platform that enables organizations to adopt a Modern Data Architecture and provide Hadoop data platform.

• With HDP for Windows, Hadoop is both simple to install and manage.

• Familiar Tools on Hadoop: The new offering enables the application of rich business intelligence (BI) tools such as Microsoft Excel, PowerPivot for Excel and Power View to pull actionable insights from not just big data but all of your enterprise data sources.
Hortonworks Data Platform (HDP) Types

**HDP Sandbox**
- Host Operating Systems: Windows 7, 8
- Virtual Machine: VirtualBox, VMWare or VMFusion

**Automated (Ambari)**
- Red Hat Enterprise Linux • CentOS • Oracle Linux • SUSE Linux Enterprise Server

**Windows**
- Windows Server 2008 R2 (64-bit) • Windows Server 2012 (64-bit)
HDP Minimum System Requirements

- Hosts:
  - A 64-bit machine with a chip that supports virtualization.
  - A BIOS that has been set to enable virtualization support.
- Host Operating Systems: Windows 7, 8
- Supported Browsers: Internet Explorer, Google Chrome, Firefox
- At least 4 GB of RAM (Divide Total RAM by half between Host and Virtual Machine)
- Virtual Machine Environments: Oracle Virtual Box - version 4.2 or later, VMware, VMware Fusion, version 5.x (For Mac)
Setting up HDP inside Virtual Machine

Welcome to VirtualBox!
The left part of this window is a list of all virtual machines on your computer. The list is empty now because you haven’t created any virtual machines yet.

In order to create a new virtual machine, press the New button in the main tool bar located at the top of the window.

You can press the F1 key to get instant help, or visit www.virtualbox.org for the latest information and news.
Setting up HDP inside Virtual Machine (Cont...)

VirtualBox - Settings

Input

Host Key: Ctrl

Auto Capture Keyboard

Select a settings category from the list on the left-hand side and move the mouse over a settings item to get more information.
Setting up HDP inside Virtual Machine (Cont...)
Setting up HDP inside Virtual Machine (Cont...)

VirtualBox currently supports importing appliances saved in the Open Virtualization Format (OVF). To continue, select the file to import below.

Open appliance...
Setting up HDP inside Virtual Machine (Cont...)
Setting up HDP inside Virtual Machine (Cont...)
Setting up HDP inside Virtual Machine

Appliance settings

These are the virtual machines contained in the appliance and the suggested settings of the imported VirtualBox machines. You can change many of the properties shown by double-clicking on the items and disable others using the check boxes below.

<table>
<thead>
<tr>
<th>Description</th>
<th>Configuration</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU</td>
<td>2</td>
</tr>
<tr>
<td>RAM</td>
<td>2048 MB</td>
</tr>
<tr>
<td>USB Controller</td>
<td>☣</td>
</tr>
<tr>
<td>Sound Card</td>
<td>☣ ICH AC97</td>
</tr>
<tr>
<td>Network Adapter (Primary)</td>
<td>☣ Intel PRO/1000 MT Desktop (82540EM)</td>
</tr>
<tr>
<td>Network Adapter (Secondary)</td>
<td>☣ Intel PRO/1000 MT Desktop (82540EM)</td>
</tr>
<tr>
<td>Hard Disk Controller</td>
<td>☣ PIX4</td>
</tr>
</tbody>
</table>

Reinitialize the MAC address of all network cards
Setting up HDP inside Virtual Machine (Cont...)
Setting up HDP inside Virtual Machine (Cont...)
HDP Console Interface

Hortonworks Sandbox 1.2
http://hortonworks.com

To initiate your Hortonworks Sandbox session, please open a browser and enter this address in the browser’s address field:
http://192.168.56.101/

Log in to this virtual machine: Linux/Windows <Alt+F5>, Mac OS X <Cmd+Alt+F5>
HDP Web Interface at 127.0.0.1:8888
What is JSON file?
What is Raw Data?
What is JSON Serde file?
How to load external data into Hive?
from windows machine
What is Dictionary File?
What is JSON file?

- **JSON** (JavaScript Object Notation) is a lightweight data-interchange format
- It is easy for humans to read and write. It is easy for machines to parse and generate
- It is based on a subset of the JavaScript Programming Language
What is Raw Data?

- Raw data is the data generated from twitter in JSON format using twitter API 1.1.
- The data has fields such as:
  - Name
  - Screen
  - Date time
  - Text
  - Hash tag
- These fields are generated when a user tweets or retweets.
- There are many other fields in the data for a particular record, which are not required for the analysis.
Really wanna see Iron Man 3 o-

o
SerDe is short for Serializer/Deserializer.

Hive uses the SerDe interface for IO.

A SerDe allows Hive to read in data from a table, and write it back to HDFS in any custom format.

Here we are using SerDe for row format.

For JSON files, Amazon has provided a JSON SerDe.
Loading external data into Hive from Windows Machine

- Raw data and JSON SerDe files are the external data
- Hive uses external data and JSON SerDe file to load external tables
- These external files are transmitted from windows to Hadoop environment, using a win SCP recommended by Hortonworks
- It is a interface to access remote system from local machine, and store files and data from an external resource
- Here remote system is hortonworks sandbox and external resource is the external data
WinSCP Screen Shots
What is Dictionary File?

- It is text file with .tsv format.
- Data is arranged in three columns
- First column is the behavior of the word. A word can have weak subject or strong subject.
- Second column contains the word.
- Third column is the polarity of the word.
- Before every word, the polarity of each word is saved i.e. positive, negative or neutral.
MAP and REDUCE functions in Hadoop
Division of Data Words
Business Intelligence Tools
How to connect HDP to MS-Excel?
Power Query via CSV
Challenges and Overcomes

Srijha Reddy Gangidi
Application Developer / Tester

Capstone Project Group 1
MAP and REDUCE functions in Hadoop

- **MapReduce** is a **programming model** for processing and generating large data sets with a **parallel, distributed** algorithm on a **cluster**.

- A MapReduce program is composed of a **Map() procedure** that performs filtering and sorting.

- A **Reduce() procedure** that performs a summary operation.

- MapReduce can take advantage of locality of data, processing it on or near the storage assets in order to reduce the distance over which it must be transmitted.
"Map" step: Each worker node applies the "map()" function to the local data, and writes the output to a temporary storage. A master node orchestrates that for redundant copies of input data, only one is processed.

"Shuffle" step: Worker nodes redistribute data based on the output keys (produced by the "map()" function), such that all data belonging to one key is located on the same worker node.

"Reduce" step: Worker nodes now process each group of output data, per key, in parallel.
The identification of subjective opinion on text data involves the classification of text into three categories: Positive, Negative and Neutral.

Positive sentiment is measured in a similar way by looking for positive words not preceded by a negation.

Similarly the negative sentiment is measured by looking for negative words.

Neutral sentiment is measured by looking for positive words preceded by a negation or vice versa.
Business Intelligence (BI) Tools

• **Business intelligence tools** are a type of application software designed to retrieve, analyze, transform and report data for **business intelligence**.

• The **tools** generally read data that have been previously stored in a data warehouse or data mart.

• The **business intelligence** (BI) represents the tools and systems that play a key role in the strategic planning process of the corporation. These systems allow a company to gather, store, access and analyze corporate data to aid in decision-making.
How to connect HDP to MS-Excel

- We use the **Power View feature in Excel 2013** to visualize the sentiment data. Other versions of Excel will work, but the visualizations will be limited to charts.

- Install the ODBC driver that matches the version of Excel you are using (32-bit or 64-bit).

- Connecting HDP to MS-Excel involves:
  - Accessing the refined sentiment data with Excel
  - Visualize the sentiment data using Excel Power View
Access the Refined Sentiment Data with Excel

• In Windows, open a new Excel workbook, then select Data > From Other Sources > From Microsoft Query.
BI – Tools in Excel (Cont..)

• On the Choose Data Source pop-up, select the Hortonworks ODBC data source you installed previously, then click OK.

• The Hortonworks ODBC driver enables you to access Hortonworks data with Excel and other Business Intelligence (BI) applications that support ODBC.
• After the connection to the Sandbox is established, the Query Wizard appears.

• Select the “tweetsbi” table in the Available tables and columns box, then click the right arrow button to add the entire “tweetsbi” table to the query. Click **Next** to continue

• ODBC configuration ERROR!
Power Query via CSV file
An alternative approach to BI – Tools in Excel

- Install power view and power query in MS Excel
- Export the table in CSV format from the web interface
- Open the table in Power Query and manage the table
- Load the manage table into excel worksheet
- Visualize it in Power view using Map view.
Power Query via CSV file – An alternative approach
Power Query via CSV file (Cont...)
### Power Query via CSV file

#### Data Preview:

<table>
<thead>
<tr>
<th>text</th>
<th>country</th>
<th>sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waiting for Iron Man 3 to start.</td>
<td>ECUADOR</td>
<td>1</td>
</tr>
<tr>
<td>Iron Man 3 &amp; Iron Man 2 for SURE ugh it was so perfect</td>
<td>UNITED STATES</td>
<td>1</td>
</tr>
<tr>
<td>Iron Man 3 was...</td>
<td>UNITED STATES</td>
<td>1</td>
</tr>
<tr>
<td>Seeing Iron Man 3 at Midnight!</td>
<td>TAIWAN</td>
<td>1</td>
</tr>
<tr>
<td>Iron man 3 was really funny xD</td>
<td>UNITED STATES</td>
<td>1</td>
</tr>
<tr>
<td>&quot;Wait&quot; @BoredPrints. Kako ppi naprijed ghobetan nih im Iron Man 3 tpa dlja vez</td>
<td>THAILAND</td>
<td>1</td>
</tr>
<tr>
<td>Iron man 3 midnight premiere...</td>
<td>UNITED STATES</td>
<td>1</td>
</tr>
<tr>
<td>RT @WhyTheBlimey: Totally forgot about Iron Man 3 tomorrow!</td>
<td>THAILAND</td>
<td>1</td>
</tr>
<tr>
<td>&quot;@ArpadWiguna Adip: Mending IRON MAN 3 yang 3D ini?&quot;</td>
<td>INDONESIA</td>
<td>1</td>
</tr>
<tr>
<td>RT @MashTriumph: I'm going scene by scene spoilers for Iron Man 3, tonight.</td>
<td>UNITED STATES</td>
<td>1</td>
</tr>
<tr>
<td>&quot;I unlocked the Marvel's Iron Man 3 Coming Soon sticker on #SendSquad <a href="http://t.co/YAMGfNoW1S">http://t.co/YAMGfNoW1S</a>.&quot;</td>
<td>UNITED STATES</td>
<td>1</td>
</tr>
<tr>
<td>If you're watching Iron Man 3, I'm jealous and I hate you.</td>
<td>UNITED STATES</td>
<td>1</td>
</tr>
<tr>
<td>Iron Man 3 was good but a comic nerd it could have been better.</td>
<td>UNITED STATES</td>
<td>0</td>
</tr>
<tr>
<td>IRON MAN 3 in theaters everywhere 5/3. that is tomorrow I loved it and</td>
<td>INDONESIA</td>
<td>2</td>
</tr>
<tr>
<td>IRON MAN 3 is an iron man...</td>
<td>UNITED STATES</td>
<td>2</td>
</tr>
<tr>
<td>&quot;Iron Man 3&quot;</td>
<td>COUNSELMAN9909</td>
<td>2</td>
</tr>
<tr>
<td>IRON MAN 3 3D Watch</td>
<td>COUNSELMAN9909</td>
<td>2</td>
</tr>
<tr>
<td>IRON MAN 3 This Weekend?</td>
<td>UNITED STATES</td>
<td>2</td>
</tr>
<tr>
<td>just got off the phone with</td>
<td>DONTLOOKATME after he watched Iron Man 3</td>
<td>UNITED STATES</td>
</tr>
<tr>
<td>Iron Man 3 premiere with my fam at the VIP @actor_bld</td>
<td>UNITED STATES</td>
<td>2</td>
</tr>
<tr>
<td>&quot;Dude, is that good? @EricIGN. Off to see Iron Man 3 for the fourth time...&quot;</td>
<td>UNITED STATES</td>
<td>0</td>
</tr>
</tbody>
</table>
Power Query via CSV file (Cont...)

![Power Query screenshot]

In the image, we see a screenshot of the Power Query interface with a CSV file loaded. The interface shows a table with the following columns: "ID", "DateTime", and "country". The table contains several rows with data corresponding to these columns. The interface also includes various editing tools and options available in Power Query, such as transforming data, adding columns, and managing queries.
Power Query via CSV file
Power Query via CSV file

In Power Query, you can import data from a CSV file and use it for further analysis. The screenshot shows a CSV file being loaded into Excel through Power Query. The data includes columns such as 'id', 'country', and 'sentiment', with rows of data from different countries.

The Power View tool is highlighted, indicating that reports can be created to make better business decisions and create beautiful, interactive reports.

The Excel sheet also shows a summary of the data, with counts and averages, and a query indicating that 89,843 rows were loaded into the workbook.
Power Query via CSV file
Map Display of Sentiment Data

- Orange : Positive
- Blue : Negative
- Red : Neutral
Challenges and Overcomes

- Encountered issues while installing Hive and Hadoop Separately
  - Switched to HortonWorks Sandbox with preinstalled Hadoop and Hive as per atlink.
- System got slow and got stuck upon installation of Hortonworks
  - Re-Divided Ram allocation equally between Windows and HDP
- Importing JSON file
  - Implemented usage of WinSCP - A file transfer software to remote machine
- Hive & MapReduce jobs not configured
  - Switched to Stable HDP 2.0 from HDP 2.2 with pre-configured Hive and MapReduce
- Currently facing the problem of ODBC Driver Configuration with Hortonworks
Sentiment Analysis using Hadoop
Sponsored By Atlink Communications Inc

Team Members: Ankur Uprit, Pinaki Ranjan Ghosh, Kiranmayi Ganti, Srijha Reddy Gangidi

Instructor: Dr. Sadegh Davari
Mentors: Dilhar De Silva, Rishita Khalathkar

Capstone Project Group 1