ISMT E-599 Capstone
Seminar in Digital Enterprise
Spring - 2016

Social Analytics

Market Insights with Social Media Analytics

Team 4:
Neelesh Pradhan
Usha Annipu
Asafali Karim
Harish Kumar

Social
Analytics

Table of Contents

1 Executive summary 2
1.1 The Client 2
1.2 ICT 2
1.3 Business Problem 2
1.4 Proposed Solution 3
2 Business Requirements 3
2.1 AS-IS and TO-BE 3
2.2 User Stories and User Roles 5
2.3 Functional Requirements 8
2.4 Non-Functional Requirements 8
2.5 Business Benefit Justification 9
2.6 Success Metrics 9
3 Technical Specification and Prototype 10
3.1 Architectural Approach 10
3.2 Software Solution Details 12
3.3 Data Design and Management 19
3.4 Administrative Support 19
3.5 Integration and Visualization Details 21
3.6 Solution Demonstration 23
3.7 Development Platform 25
3.8 System Metrics 25
3.9 Vendor Selection Criteria 26
4 Implementation Plan 27
4.1 Solution Development 27
4.2 Solution Deployment 30
4.3 Operationalization of the Solution 34
4.4 User Enablement 35
4.5 Project Risks 36
4.6 Success Metrics 36
5 APPENDICES 39
Appendix : A 39
Appendix : B 39
6 References 39

1 EXECUTIVE SUMMARY

1.1 THE CLIENT

Gloco is a multi-national, medical manufacturing and services company headquartered in Cambridge, MA. The company has nearly thirty thousand employees and a network of manufacturing services around the world. Gloco has expanded its footprint from manufacturing hospital grade medical equipment to consumer grade devices. For instance, Gloco manufactures consumer devices such as baby monitors, blood glucose monitors, blood pressure machines and wearable consumer devices. The company has now decided to enter into the digital services realm and intends to be a world leader in offering digital solutions to the medical industry.

1.2 ICT

As a result of Gloco's digital technology initiatives, ICT, the Information and Communications Technology arm of Gloco, has only risen in prominence. ICT has a history of creating quality digital applications in-house, as well as using third party vendors in its efforts to build custom applications according to industry needs. The ICT has observed the recent rise and evolution of business analytics and the power it holds to help businesses transform their organizations. The ICT therefore, has decided to delve into building visualization software and enterprise-wide high performance rich analytical solutions.

1.3 BUSINESS PROBLEM

Today's digitized world has opened up new avenues for expanding the brand visibility of an organization. The ubiquity of social media has made the voice of the customer all pervasive. As such, it has become a business imperative for organizations to be able to listen to and harness the voice of the consumer. The social analytics umbrella encompasses the collecting, filtering, mining, classifying, measuring and analysis of data obtained from social media to improve business decisions. According to a Gartner article, in the period from 2014 to 2015, there has been an 87% increase in companies that view social analytics as a part of their broader business strategy [1].
Gloco currently finds itself on the wrong end of this trend. Not having an automated social analytics strategy in place, Gloco is unable to gauge the social market penetration of its medical products. Not being able to track customer sentiments in real time, Gloco is unable to measure the reputation of its products in the social media sphere. As a result, Gloco is steadily losing its competitive edge in the marketplace. Customer experience, customer service satisfaction, brand health, brand reputation, the ability to identify negative opinions etc., are all critical parameters to be able to make more efficacious business decisions. To that end, Gloco has identified the following business goals where social media analytics can play an essential role:

- Assessing customer experience with the products.
- Assessing customer service experience.

- Assessing brand health and reputation.
- Assessing customer proclivity to buy a particular product by region or demographic.
- Identifying effective social media avenues and marketing channels to reach out to the consumer.
- Raising brand awareness, creating a brand identity and positive brand association.
- Improving interaction with customers and identifying ways to engage directly with them.
- Identifying influencers and their opinions about Gloco products.

1.4 PROPOSED SOLUTION

To address these business goals, Gloco will begin an exploration of the latest social media analytics technology solutions. Gloco will evaluate, implement and customize social media monitoring and analytic tools that provide functional and actionable customer intelligence. The social analytics solution should provide:

- Interactive visualization with dashboards and charts.
- Real-time insights with response capabilities.
- Monitoring tools that can capture user sentiments.
- Conversation monitoring that listens in on prospective customer conversations and identifies marketing and sales opportunities.
- Industry competitor monitoring capabilities.

Gloco will also employ custom text mining techniques to better tailor the data for its use. Gloco will acquire data from multiple social networks, industry specific communities, forums and blogs and devise custom data mining algorithms that will mine the unstructured data and extract usable customer intelligence. Dictionary based sentiment analysis techniques as well as machine learning techniques will aid in the information analysis of the obtained data. The usable data can then be correlated with the Gloco CRM to further enhance its value. Metrics such as net sentiment, customer satisfaction, leads generated etc., will enable Gloco to gauge the success of the social analytics solution and help meet its business goals.

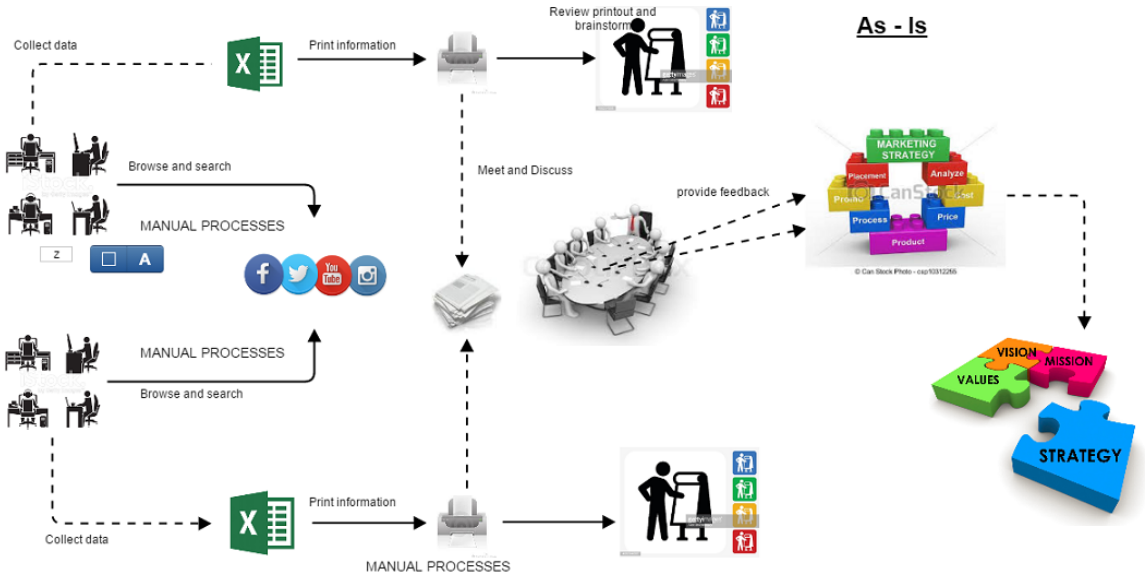
2 BUSINESS REQUIREMENTS

The business requirements section describes the business context and establishes the business requirements for the Gloco social media solution. The section documents the functional and non-functional requirements, the AS-IS and TO-BE processes, the user stories, the acceptance criteria and identifies the user roles, user functions and user workflows. The section concludes with a business benefit justification for the social analytics solution.

2.1 AS-IS AND TO-BE

The AS-IS process:
The current business process for social media analytics is a manual, time-consuming process where the users collect social media conversations about Gloco and its products in

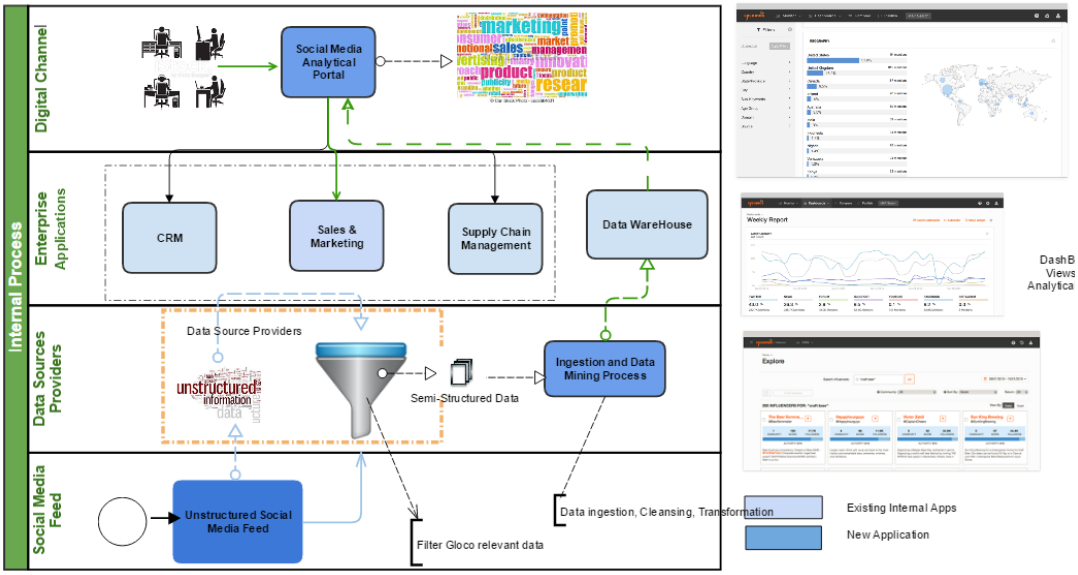
spreadsheets. Meetings and brainstorming sessions between social media team members are held to evaluate the conversations and a consolidated report is provided to sales and marketing teams. This manual process needs to be redesigned to help Gloco formulate a more efficacious social media strategy.



TO-BE process:

- The unstructured data feeds from social networks Twitter and Facebook will be automatically filtered based on configurable rules. The semi-structured data is ingested, cleansed and transformed by the data mining process.
- The structured data from the data mining process will be fed into another process that performs sentiment analysis to classify the data as positive, negative and neutral.
- The classified social media data will be accessible through a new analytical portal, which provides capabilities to perform social media analysis much more efficiently by using interactive dashboards and graphical representation of large amounts of data.
- The social analytical portal provides capability to retrieve customer and sales information for a particular conversation of interest by integrating the analytical portal with internal enterprise applications.
- Capturing this quantitative data originating from unstructured social comments will enable Gloco to monitor brand perception, market trends and other important business interests.

To-Be Process Diagram



2.2 USER STORIES AND USER ROLES

Solution User Role Identification

Detailed below are user roles and a description of what filters they can define, what sentiment reports they can access, and what actions they can undertake. We identify Business User Roles and Administrator User Roles below.

Business User Roles:

Marketing Strategist

This user role is involved with managing the overall market strategy of the brand in the social media space. This strategist will be able to filter the social media data by particular brand names and products and assess the sentiment associated with them. Based on this, the strategist can create or tailor the broader marketing strategies for Gloco.

FILTERS	REPORTS	ACTIONS
• Define brand names, products, and people to filter and measure.	• Sentiment • Brand health • Brand awareness	• Guide marketing strategies, and conceive brand/product-messaging campaigns based on brand health and consumer sentiment.

Product Strategist

This user role will be involved with defining and determining the overall product strategy. This strategist will be able to filter the social media data by particular product names and assess the sentiment associated with them. Based on this, the strategist can tailor the product strategy for Gloco.

FILTERS	REPORTS	ACTIONS
• Define products, product features to filter by.	• Product sentiment • Product feature sentiment	• Determine product and product features, enhancements etc. based on consumer sentiment and feedback.

Sales Strategist
This user role will be involved in driving sales for Gloco products.

FILTERS	REPORTS	ACTIONS
• Brand names. • Trending topics	• Conversations of customers	• Identify needs/wants of customers and reach out to them. • Join conversations and spread product awareness

Customer Strategist
This user role will handle the customer questions, complaints and service needs.

FILTERS	REPORTS	ACTIONS
• Brand names • Products • People	• Monitor conversations of customers. • Influencer conversations.	• Respond to customers with complaints or service needs. • Track influencers, assess their sentiment and engage with them.

Administrator User Roles:

Super administrator

ACTIONS
• Create and administer users • Create and delete social media analytics projects • Configure the projects and import/export project configuration data.

Data specialist

ACTIONS
• Verify the data in HDFS using tools such as Solr search and HCatalog. Monitor and manage the Cloudera cluster using Cloudera manager admin console. Configure and update dashboards.

User Stories

The use cases for the proposed social media analytics solution are listed below.

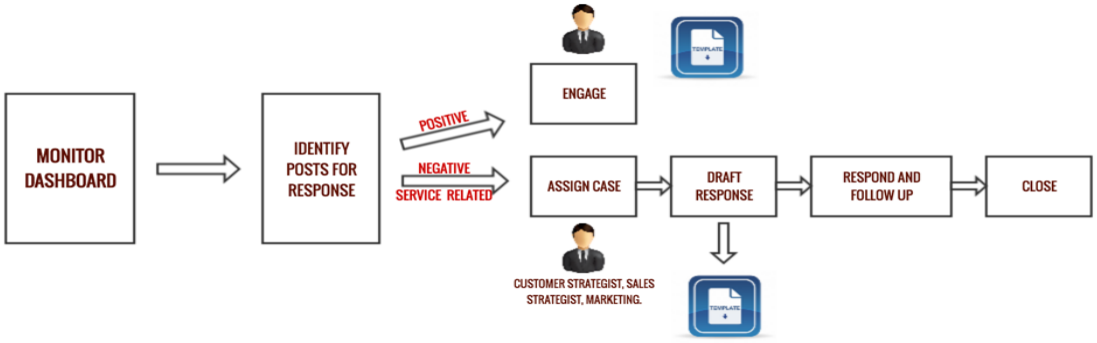
#	Use Case	Acceptance Criteria
1	As a Marketing Strategist at Gloco, I need to view a heat map of Gloco's consumer sentiments.	• The system should present a real time dashboard with a world map as background and display the overall sentiments in green, amber and red colors for a region or country. • User should be able to locate regions where company reputation is good as well as regions where reputation needs improvement. • User should be able to filter by region, country, product and social media channel. • User should be able to drill down or zoom into a particular geo-location. • Colors should represent aggregate sentiments at each drill down level, green for positive sentiment, amber for neutral and red for negative sentiment. • User should be able to export the dashboard as a PDF file.
2	As a Product Strategist at Gloco, I need to access sentiment analysis trend reports.	• The system should present a dashboard with charts showing the sentiment trend over a period of time - by months, weeks or days. • The x-axis will be the time period and y-axis will be the number of posts. • User should be able to filter by region, product, sentiment and social media channel. • Different colors should be used for different social media channels. • User should be able to export the dashboard as a PDF file.
3	As a Customer Strategist at Gloco, I need to track social media conversations about Gloco, its products, and its competitors.	• The system should present a dashboard to track conversations of customers as well as influencers. • The system also allows the user to search by keywords and find top influencers. • It should be possible to sort by a score assigned to the influencers. • User should be able filter by region, product, and social media channel. • User should be able to export the dashboard as a PDF file.

User Engagement Workflow

Customer Strategist Role Workflow Example:

Following is an instance of how a Customer Strategist role will use the system:

- Monitor dashboard and identify posts for response.
- If a positive post, engage with the user. Follow set response template if applicable.
- If negative/service related post, assign the case to the appropriate role.
- Draft response after evaluation. Use set response template if applicable.
- Respond and follow up until issue resolved.
- Close the case.



2.3 FUNCTIONAL REQUIREMENTS

In this section we identify the specific functional features the solution should incorporate. The Gloco social media solution should:

- Enable API access to data feeds from major social media networks such as Facebook and Twitter as well as relevant sources from the wider blogosphere.
- Provide a filtering capability to extract the social media data by company, competitor and product.
- Perform a sentiment classification of the social media data by identifying the polarity (positive, negative or neutral) of each social post.
- Provide dashboards that enable the visualization of the sentiment classification, products mentioned, customer information and social media influencers.
- Provide a user administration console to create and manage user roles, user permissions, as well as a UI to create projects based on different filtering configurations.
- Allow for integration with internal enterprise applications to match social media data with customer and sales information.

2.4 NON-FUNCTIONAL REQUIREMENTS

The Gloco social media solution should incorporate the following non-functional requirements:

- The social analytics solution will facilitate security and access management by enabling Single Sign-On (SSO) and integrate with the Gloco corporate portal.
- The system will be highly responsive and able to handle 100 concurrent users with a 30 seconds response time.
- The solution will be highly scalable.
- Maintain interoperability with the interfaces.
- The solution will ensure security standards are implemented.
- The solution will satisfy any audit requirements.
- The solution will adhere to Gloco's standard technology stack.

- Usability documentation and training on how to operate the solution will be available for Gloco users.

2.5 BUSINESS BENEFIT JUSTIFICATION

- The social media analytics solution will help Gloco regain its competitive edge by enabling it to chart and execute an effective social media strategy.
- The social media analytics solution will help Gloco grow revenue and improve financial performance by employing a strategy to increase business through the digital marketing channels.
- The social media analytics solution will enable Gloco to automate and improve the operational efficiency of tracking online conversations.

Project Projected Costs

	Year 0	Annual
Software & licenses	$1M	$700K
Development	$2.25M	
Hosting	$500K	$750K
Software maintenance		$500K
Training	$250K	
Data feeds	$500K	$600K
Total	**$4.5M**	**$2.55M**

	Year 0	Year 1	Year 2
Units Sold	600,000	606,000	618,120
Unit price	$1,500	$1,501	$1,500
% increase in units sold		1%	2.0%
Revenue Increase	0	$9,006,000	$18,180,000
Solution cost	$4,500,000	$2,055,000	$2,055,000
ROI %	-100%	338%	785%

2.6 SUCCESS METRICS

Business Goals	Use Cases	Functional Requirements	Success Metrics
Raising brand awareness, health and reputation by social media outreach and marketing channels/campaigns to reach out to the consumer.	Use case 1: Access a sentiment analysis heat map and conduct a social media outreach to improve brand health by region.	Perform a sentiment classification of the social media data by identifying the polarity.	5% Increase in brand awareness and satisfaction. 2% reduction in Marketing expenditures in campaign outreach because of improved health of brand.

Increasing customer proclivity to buy a particular product by region or demographic.	Use case 2: Access sentiment analysis trend report for a brand/product.	Provide capability to extract the social media data by company, competitor and product and perform a sentiment classification.	Increase in product sales by 5%.
Improve overall Customer experience	Use case 3: Track social media conversations about Gloco, its competitors and about Gloco products.	Provide a capability to extract the social media data by company, competitor and product.	Based on the number of positive comment improve the customer experience by 5%
Improve customer service experience with the product	Use case 3: Track and identify the customer social media conversations about Gloco product defects/complaints	Provide tools to find out the customer experience and track the conversation, identify the customer is an existing product user and follow up.	10% reduction in customer complaints. 5% improvement in customer services.
Improve interaction with customers and identify ways to engage directly with them.	Use case 3: Track mechanism to follow up on new prospect and lead opportunity.	Integrate with internal CRM, SCM to match data with customer and sales information. Provide dashboard for Marketing users and sales user to follow up on leads and prospects.	5% increase in lead generation and 2% increase in opportunity lead to sales closeouts. 3% increase in prospects and 1% in revenue.

3 TECHNICAL SPECIFICATION AND PROTOTYPE

3.1 ARCHITECTURAL APPROACH

In order to meet its business goals, Gloco will implement a custom social media analytics solution that can provide functional and actionable customer intelligence. To that end, we seek to formulate a unified information architecture that combines varied data sources into a robust analytics model.

Architectural Principles

Architectural principles define the architectural design approach that will drive the analytics solution. The social media analytics solution for Gloco will incorporate the following architectural principles:

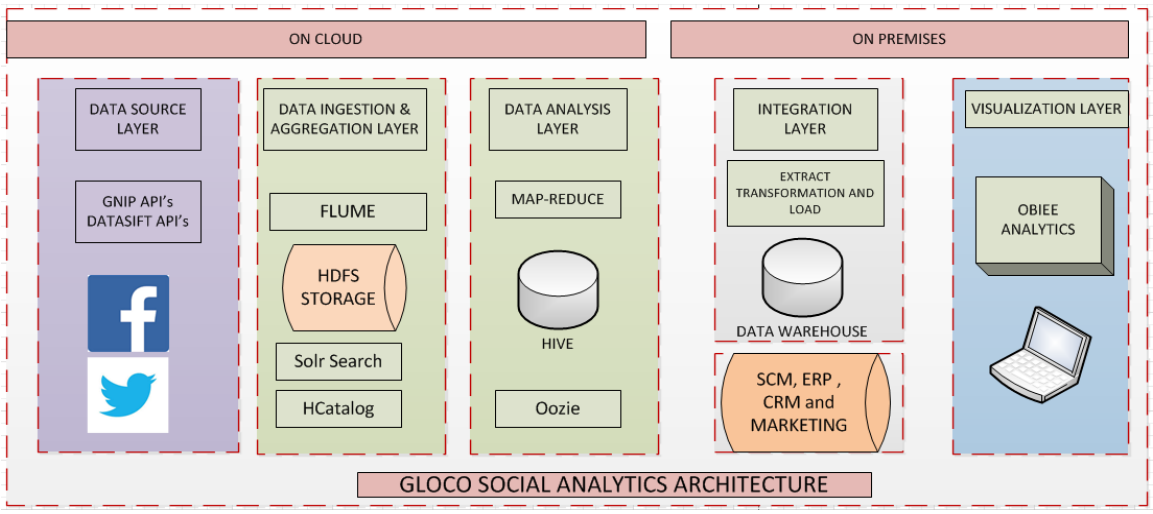
- All data sources that help Gloco meet its business goals will be supported.
- The architecture built will be flexible enough to support multiple sources of real-time high volume data.

- The architecture will provide mechanisms to cleanse, format, aggregate and structure the data obtained from multiple sources.
- Data will be readied to support analytic object models.
- The solution will support integration with existing enterprise systems in order to leverage all available information.
- The solution will present intuitive and interactive visualization.
- The analytics solution will make it easy to turn insights into actions and meet the business goals.

Logical layers

We take a layered architectural approach to organizing specific components of the analytics solution. The Gloco social media analytics solution will be comprised of the following logical layers:

1. Data Sources
2. Data Ingestion and aggregation layer
3. Analysis layer
4. Integration layer
5. Visualization layer



Data sources

DataSift and GNIP are two leading enterprise data platforms that provide APIs to access social media data. Both companies provide streaming APIs and information in near real-time. These data sources will provide us with the following data:

- DataSift will provide Gloco filtered access to Facebook topic data via Datasift's PYLON API.
- Gnip will provide Gloco filtered access to the twitter firehose via Gnip PowerTrack streaming API.

***Format**:* The data will be accessed in a semi-structured JSON format.
***Volume**:* Thousands of lines per second will be accessed.
***Collection Point**:* The data from the Datasift and Gnip APIs will be accessed via Apache Flume.
***Destination**:* Apache Flume will push the data to a Hadoop cluster in the cloud.

Data ingestion and aggregation layer
In this layer the social media data from DataSift and Gnip API's will be accessed via Apache Flume. Apache Flume is a data ingestion system that collects and aggregates large amount of streaming data into the HDFS (Hadoop Distributed File System). HCatalog is a metadata and table management system for HDFS. HCatalog contains the structural information of the table that can be used to build a relational view of the HDFS data. The raw JSON data from the HDFS will be projected into a relational format using HCatalog and a Hive script.

Analysis layer
This layer will analyze the data by using Apache Hive. Hive is the de facto standard for SQL interaction with Hadoop data. Hive queries, in combination with the data analysis algorithms, will be executed in this layer to perform data analytics. This encompasses Gloco use cases such as sentiment analysis, trend analysis and sales data analysis. MapReduce techniques will aid in processing the raw data. The analyzed data will then be pushed to a data warehouse.

Integration layer
The analyzed social media data will be integrated with Gloco's enterprise systems. Existing On-Line-Transaction-Processing (OLTP) data like the ERP, CRM, SCM and marketing data will be cross-referenced with the analyzed social media data. The corresponding data will be pulled (ETL) into the On-Line-Analytical Processing (OLAP) data warehouse.

Visualization layer
This layer will present the output provided by the analysis layer. Business Intelligence (BI) tools and visualization applications will animate rich, interactive dashboards, graphs and charts. The visualization tools will help the users identify sentiments, patterns, trends and relationships. Oracle Business Intelligence Enterprise Edition (OBIEE) suite of applications will be employed as the visualization solution.

3.2 SOFTWARE SOLUTION DETAILS

This section will describe the Gloco social media analytics solution in detail. As described in the architectural approach, we shall employ Gnip and DataSift as the primary providers for Twitter and Facebook data.

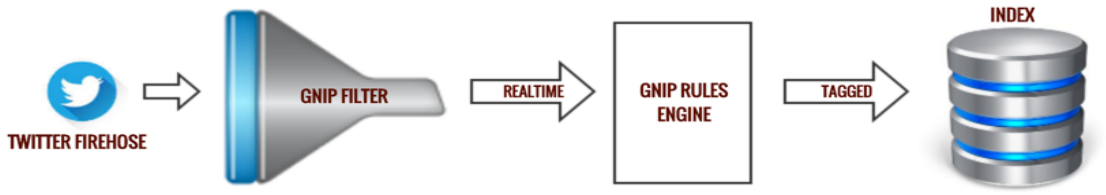
Gnip product **PowerTrak** offers real time enterprise access to twitter data. Twitter data comprises of data related to each individual tweet including the metadata, geolocation, images and mentions. Also included are activities associated with the tweet such as whether the tweet is a retweet, a deleted tweet, or an undeleted tweet. User stats such as favorites count, follower count, etc. are also made available.

DataSift product **PYLON** provides access to Facebook topic data. Facebook Topic data consists of stories, which are posts made by an individual user on their own timeline as well as engagements, which are interactions on a story posted by someone else. The story data can contain the text of the post, links shared, videos, images, albums, reshares and any long form content that is available. The engagements data can contain the number of likes on a story, comments and number of shares. Demographic data that includes the type, age, gender and location of the author is also made available.

Filtering Process

Let's look at the filtering process that the Gloco social analytics solution will employ to get the relevant social media data. The overall process is very similar for both Gnip and DataSift with some minor technical differences.

Gnip filtering process:

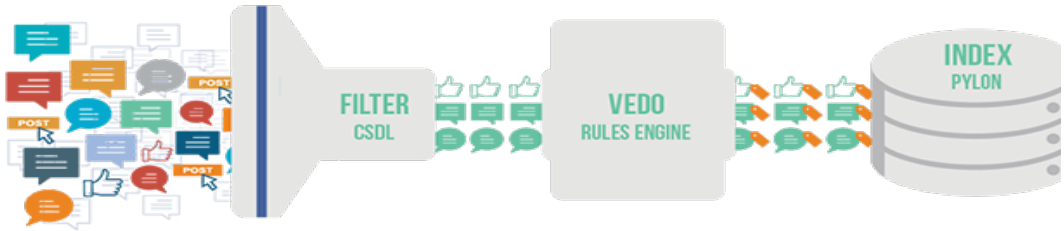


The PowerTrak API provides the ability for a consumer to filter the available social media data and access only the data that is needed. Filtering rules can be set up to enable the enterprise to filter the data according to its needs. PowerTrack stipulates the use of its own filtering language, which has certain syntax, operations and keywords that allow the consumer to formulate their own filtering rules. Boolean logic is used to group multiple keywords. For instance, to access the twitter data for Gloco and a competitor, the OR logic type is used:

```
String rule1_value = "(Gloco OR GlocoCompetitor) (GlocoBabyMonitor)";
String rule2_value = "(Gloco OR GlocoCompetitor) (GlocoDiabetesMonitor)";
String rule3_value = "(Gloco OR GlocoCompetitor) (GlocoMRIMachine)";
// Rule syntax to add.
        String rule_tag1 = "GlocoBabyMonitor";
        String rule_tag2 = "GlocoDiabetesMonitor";
        String rule_tag3 = "GlocoMRIMachine";
// Tag for rule. Tags are optional, but      recommended!
String query1 =
String.format("{\"rules\":[{\"value\":\"%s\",\"tag\":\"%s\"}]}",rule_value, rule_tag1);
        String query2 =
String.format("{\"rules\":[{\"value\":\"%s\",\"tag\":\"%s\"}]}",rule_value, rule_tag2);
```

In the above code we are filtering twitter discussion that contain mentions of Gloco and a Gloco competitor. We are further looking for mentions of a particular Gloco product. The whitespace between `(Gloco OR GlocoCompetitor)` and `(GlocoBabyMonitor)"` represents an AND operator which denotes that we are also matching on a particular Gloco product. Since PowerTrack expects rules to be formatted as JSON, we construct them in this manner: `{"value":"(Gloco OR GlocoCompetitor) tag:GlocoBabyMonitor }`
Notice we are tagging the results by a particular product so that we can easily reference and sort the results in our business logic.

DataSift filtering process:



DataSift stipulates a two-stage filter and classification process for selecting social media data. The DataSift language for filtering unstructured data is called CSDL.
In the first stage, a filter is composed to look for posts with particular brand mentions. For instance, we filter by Gloco and a Gloco competitor in the following code:

```
// Compile filter looking for mentions of brands
       String csdl = "interaction.content contains_any \"Gloco, GlocoCompetitor \"";
String stream = datasift.compile(csdl).sync();
```

The property interaction.content in the code above encapsulates the text of a Facebook Topic. To add tags you can further classify the data thusly:

Code to classify and add tags:

```
String csdl ="tag.brand \"Gloco\" { interaction.content     contains \"Gloco\" } " +
"tag.brand \"GlocoCompetitor\" { interaction.content contains \"GlocoCompetitor\" } "+
       tag.product \ "return { " +
       "interaction.content contains_any \"GlocoBabyMonitor,
GlocoDiabetsMonitor, GlocoMRIMachine\" " +"}";
```

As illustrated in the code, the data is tagged by brand and products names.

Data Output

The data is streamed in JSON format by both solutions. We consider the Twitter data streamed by Gnip in this section. The format of the output can be broadly characterized as containing the user information, the individual tweet information and the activity information. The user information refers to the originator of the tweet, while the activity identifies whether the tweet is a retweet, deleted tweet or a quoted tweet. Below we identify and document the JSON objects that are part of a Gnip output stream.

id: Each tweet is tagged with a unique identifier so as to make it easy to store and identify it:

```
"id": "tag:GlocoBabyMonitor,2016:347769243409977344"
```

actor: The actor object encapsulates a user and contains all the pertinent user information. This includes the user id, the name of the individual tweeter, number of followers, number tweets and number of favorites.

```
"actor":
 {
    "objectType": "person",
    "id": "id:twitter.com:277184168",
    "link": "http:\/\/www.twitter.com\/KidCodo",
    "displayName": "Zach Codo",
    "postedTime": "2011-04-04T21:31:20.000Z",
```

```
    "image": "https:\/\/si0.twimg.com\/profile_images/_normal.jpeg",
    "summary": null,
    "friendsCount": 64,
    "followersCount": 207,
    "listedCount": 1,
    "statusesCount": 11207,
    "twitterTimeZone": "Central Time (US & Canada)",
    "verified": false,
    "utcOffset": "-21600",
    "preferredUsername": "KidCodo",
    "languages":
    [
       "en"
    ],
 }
```


location: The location object captures the location of origin of the tweet. The geo object pinpoints the coordinates of that location.


```
"location":
    {
    "objectType": "place",
    "displayName": "Cambridge, MA",
    "name": "Cambridge",
    "country_code": "United States",
    "twitter_country_code": "US",
    "link": "https://api.twitter.com/1.1/geo/id/fd70c22040963ac7.json",
    "geo": {
        "type": "Polygon",
        "coordinates":
        [
            [
                [
                    -105.3017759,
                    39.953552 ] "twitter_place_type": "city"}
```


body: The body object encapsulates actual text of the tweet.


```
"body": "Have to say the Gloco DRX series baby monitor is the best baby monitor on the
market. Hearing mama mama from the baby monitor brings a smile to my face every morning”
```


verb: The verb object denotes if the tweet is an original post, a retweet or a deleted tweet. The values are “post” (original post) , “share” (retweet) and “delete (deleted tweet)”. For instance: `“verb”: “share”`

The above data stream documentation only lists the main sections of the twitter stream. There is a lot more metadata and multiple subsections that are provided in the data stream. For instance, the “matching-rules” section contains tag and value sub objects that tell the user which tags and filter values the tweet matched. In our case we filter on the Gloco/competitor brand and tag by a particular Gloco product.

Filtering Administration

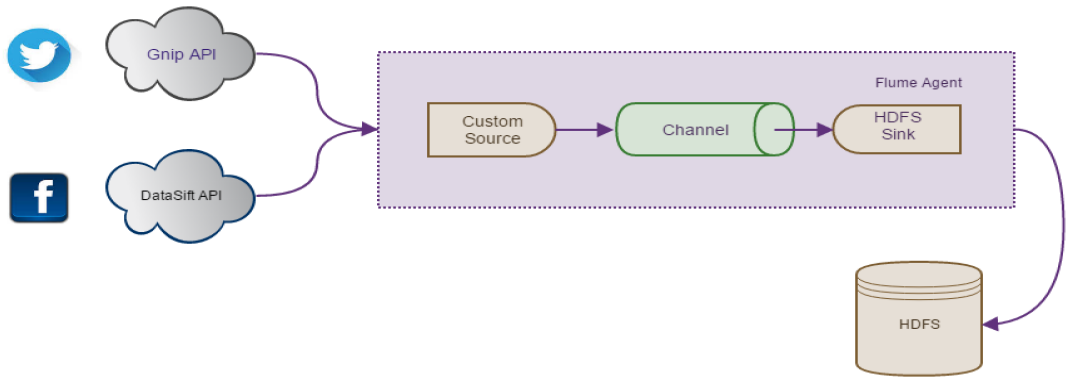
The Gloco social media analytics solution will provide a user interface console for adding filtering rules. The filtering UI screen will allow the users to configure filtering rules without having to write code. Utilizing Boolean operators, clauses and keywords, the user can easily set the filtration rules for each project. The users can set keywords for particular brands, products, and features. The UI will translate the phrases into the JSON needed to query the

data sources. The Gloco social media analytics solution will have the ability to create several projects with different filtering criteria.

Data Ingestion

The stream of semi-structured JSON data from Gnip and DataSift needs to be brought into the HDFS (Hadoop Distributed File System). The Hadoop ecosystem will then provide tools such as Hive and Oozie to analyze, query and partition the social media data. The Gloco social media analytics solution will employ components of the Hadoop ecosystem available in CDH, which is Cloudera's open-source distribution of Apache Hadoop. We will use Apache Flume to bring data into HDFS using CDH.

Data Ingestion using Apache Flume

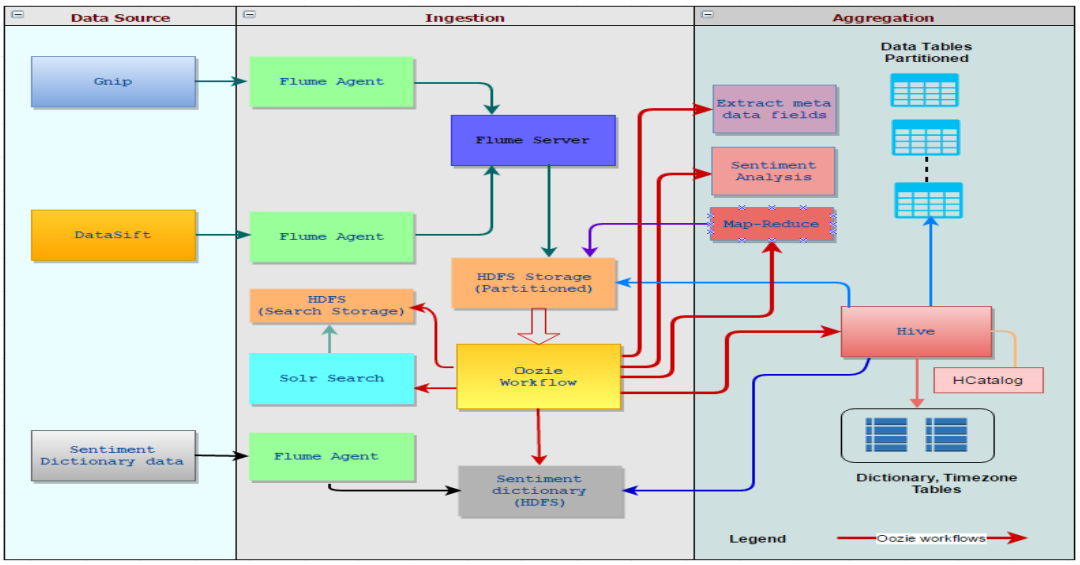


Apache Flume is described as a "distributed and reliable service for efficiently collecting, aggregating, and moving large amounts of data" by the flume documentation at flume.apache.org. The primary structure of Flume is composed of three main components, namely, sources, channels and sinks. A Flume agent is a Java Virtual Machine (JVM) process which hosts these three (source, channel, sink) pluggable components. The unit of data that flows through the source, channel and sink is called as an event.
The Gloco social analytics solution will collect the data from the Gnip and DataSift APIs and sink it into a HDFS. The first step of this process is to create a custom source. This custom source will connect to the data streamed from the Gnip and DataSift APIs, convert the discrete data into events and then push the data into the Flume channel. The custom source will employ the Gnip4j open source library to access and process activities (tweets) from the Gnip API.

A Flume channel acts as a message queue that allows for the source and sink to operate at different rates. The sink is the final component of the data streaming process inside of Flume. The Flume sink accesses the events (dataflow) from the channel and forwards them to the HDFS. The Gloco social analytics solution uses a HDFS sink, which is configured to forward the data to a preformatted location in the HDFS. As evident, all of the components described in this process reside in the cloud.

Data Analysis Process

As we've seen the data from Gnip and Datasift is ingested into Hadoop HDFS using Flume agents and stored in separate HDFS stores. Apache Oozie will orchestrate the subsequent data analysis processes. This process is depicted in the following diagram:

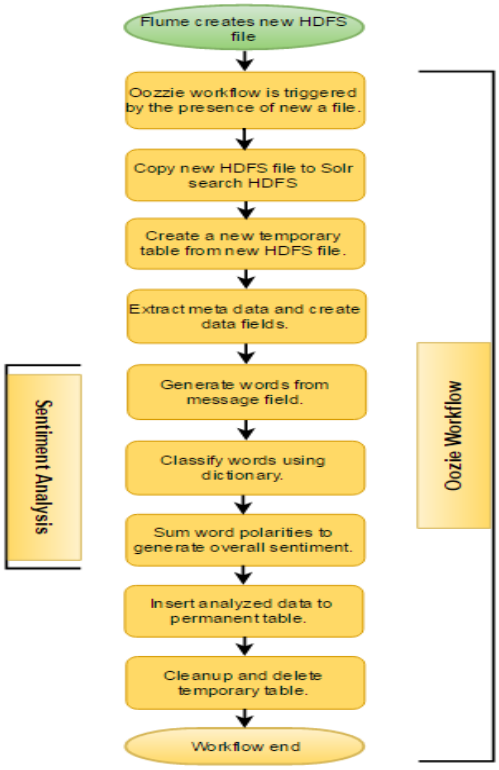


Apache Oozie

Apache Oozie is a scalable and reliable workflow scheduler to manage Hadoop jobs such as Map Reduce, Hive, Java programs and scripts. Oozie will be used to create workflows and orchestrate the processes in ingestion and aggregation layer. The JSON formatted data from Gnip or Datasift will be stored to HDFS as a new file. Oozie will be configured to trigger a "Coordinator job", when a new HDFS file is created. The process is displayed below:

Oozie workflow

When new data is received, an Oozie workflow is triggered. Oozie workflow uses Hive to create a temporary table from the new data. The workflow will have jobs to add country and time zone data based on static data tables.



Oozie will also run jobs to extract the following data elements and store them as table data fields. The fields are:
source_id, source, author_name, topic, geolocation, address, keywords, product, company, tag_names, competitor_yes_no, enagement_type , no_of_influencers, no_of_followers, language, link.
Details of the above fields are mentioned in data design section below.
Next the workflow uses the sentiment dictionary table to generate sentiment polarity for the new messages. Words are generated from the message field. Dictionary table is used to classify each word as positive, negative or neutral. The polarity for all words of the message are summed to generate overall sentiment of a message. Finally newly analyzed data will be inserted to a permanent Hive table.

The following jobs will be defined in Oozie workflow:

Oozie jobs

#	Name	Frequency	Type	Description
1	Solr_scheduler	Every new stream of data	Java MR Job	Copy HDFS files from source to solr search HDFS
2	Create_dictionary_job	When new corpus is available	Hive script	Creates dictionary for sentiment analysis and meta data tables to correlate data from social media.
3	Create temp_table	Every new stream of data	Hive script	Create a new Hive table for analysis from HDFS file.
4	Populate_meta_data	Every new stream of data	Java MR job	Extract source, keywords, products, company name etc. from message and update table fields.
5	Generate_words	Every new stream of data	Hive script	Create words from message
6	Classify_words	Every new stream of data	Java MR job	Create word polarity using the sentiment dictionary
7	Sentiment_aggregation	Every new stream of data	Java MR job	Aggregate sentiment polarity at message level
8	Insert_to_all_data	Every new stream of data	Hive script	Insert the newly analyzed data to all data Hive table
9	Cleanup_temp_table	Every new stream of data	Hive script	Delete the newly analyzed from temp table
10	Cleanup_HDFS_files	Weekly	Java MR job	Delete data older than 1 year
11	Cleanup_Hive_data	Weekly	Hive script	Delete data older than 1 year
12	Solr_cleanup_index	Weekly	Java MR Job	Delete indexed data older than 3 months

Sentiment Analysis

Sentiment analysis will be done using a "lexicon-based" approach with a dictionary suited for consumer products. In the lexicon-based approach the overall sentiment polarity of a message or post is the sum of polarities of individual words in the message.
The dictionary will have following structure:

Type	Length	Word	Part of Speech (POS)	Stemmed	Polarity
strongsubj	1	admirable	adj	n	positive
strongsubj	1	accolade	noun	n	positive
weaksubj	1	abandon	verb	y	negative
strongsubj	1	abase	verb	y	negative
weaksubj	1	all-time	adj	n	neutral
weaksubj	1	apparent	anypos	y	neutral
weaksubj	1	above	anypos	n	positive

The above is a sample of the Multi-Perspective Question Answering (MPQA) subjectivity dictionary retrieved from http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

Dictionary data will be retrieved using Flume and stored in an HDFS file. An Oozie initiated, Hive job will create the dictionary table. This table will be used for sentiment analysis. This process of creating dictionary table is repeated whenever new dictionary data is available. After a new social media message is received, and saved in a Hive table, sentiment analysis jobs are run. Sentiment analysis jobs will use the dictionary table data to generate sentiment polarity of the words in a message. Numeric values will be used to represent

sentiment polarities, 1 for positive, -1 for negative and 0 for neutral sentiments. Next, sentiment polarity for the words in a message will be summed to calculate the overall sentiment of the message and stored in a table. Sentiment value > 0 will classify the message a positive, value < 0 will indicate a negative and value = 0 will indicate the message as neutral sentiment.

3.3 DATA DESIGN AND MANAGEMENT

The data from the Hive table will be pulled into a data warehouse for further analysis and visualization. The final Hive table structure will be as follows:

#	Field	Data Type	Description
1	Uid	bigint	Unique identifier for a record. Auto generated
2	source_id	string	Original tweet id or message id from source media.
3	Source	string	Social media name like Twitter, Facebook
4	timestamp	timestamp	Timestamp of the message
5	author_name	string	Author of the message
6	user_id	string	Social media user id
7	Message	string	Message (Tweet or text posted)
8	Topic	string	Topic of the message
9	sentiment	tinyint	Sum
10	geolocation	string	Geo location
11	Address	string	Address
12	keywords	string	Keywords from message
13	Product	string	Products mentioned in message
14	company	string	Company name extracted from message
15	tag_names	string	Tags extracted from JSON data
16	competitor_yes_no	boolean	Indicate company mentioned in message is competitor or not
17	engagement_type	string	Social media network where the message is posted
18	no_of_influencers	Int	Calculated field based on number of retweets or followers
19	no_of_followers	Int	Number of followers or connections
20	Score	Int	Calculated score value(0 – 100) based on retweets, followers and shares
21	Link	string	Link to the actual tweet or message
22	Language	string	Language of tweet or post

Oozie will be used to partition the final Hive table based on a date range criteria. New data arriving from Flume will be stored in a temporary table. After processing, the data will be moved to a permanent data table, partitioned by date range. Date range based partitioning will provide scalability with data growth and better performance for queries or searches. Oozie jobs will periodically clean both HDFS files and Hive data.

3.4 ADMINISTRATIVE SUPPORT

Gloco needs to verify the unstructured social media data for audit and maintenance purposes. Providing access to raw social media data helps users verify that the filters are working properly and help Gloco improve over time. The data also helps identify false positives of the classification. Gloco's analytics project will integrate with Cloudera search

engine powered by Solr, to provide administrative support and first-hand view of raw social media data.

Solr is reliable, scalable and fault tolerant open-source enterprise search engine providing distributed indexing, replication and load-balanced querying capabilities. SolrCloud provides distributed and near real-time indexing and advanced full text search capabilities and standard open interfaces like XML, JSON and HTTP. Solr provides comprehensive HTML administration interfaces for users, which will be used to search on the indexed documents. The job to create Solr index documents, will be configured as part of Oozie work flow. Raw JSON file for each activity on HDFS will be processed by Solr search engine to create index files of the documents for searching. Once the Solr search engine is configured on the cloud servers, start the instances in Solr cloud mode and create core with name *glocosocialmedia.*

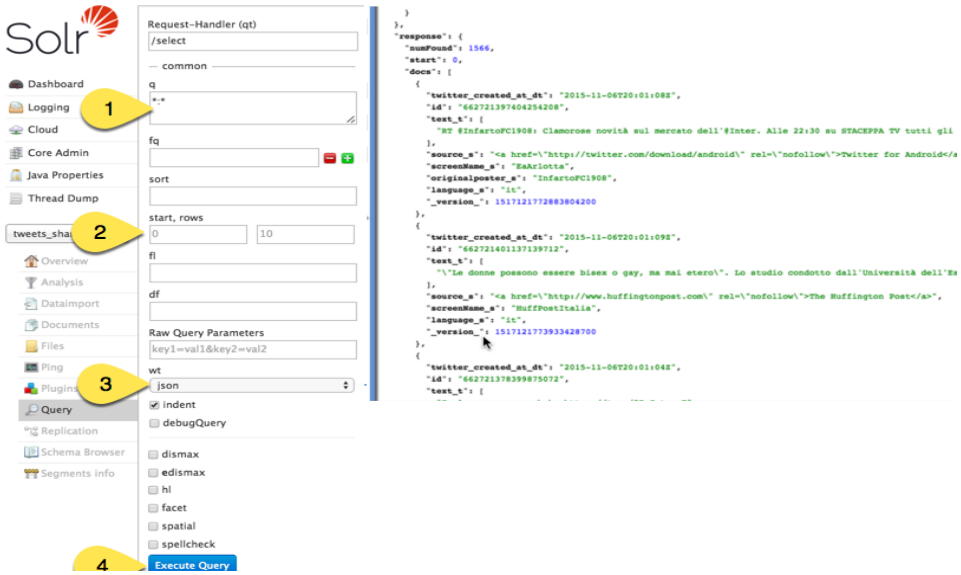
JSON schema of the message is complex and needs to be arranged into a schema type of format that users can search on individual fields. This can be achieved by including some parameters with the update request to Solr. These externally defined parameters provide information to the indexer on how to split a single JSON file into multiple Solr documents and how to map fields to Solr's schema. Below is the curl command that uses 'split api' to split the json into multiple searchable indexed documents.

```
curl 'http://gloco.com:8983/solr/glocosocialmedia/update/json/docs''?split=/''&f=author_name:/actor/displayName''&f=timest
amp:/postedTime''&f=message:/body''&f=source_id:/id''&f=no_of_followers:/actor/followersCount''&f=tag_names:/tag' -H 'Conte
nt-type:application/json' --data-binary @/gloco/socialmedia_data/2016/03/15/2016_03_15_01_02_activity.json
```

The Solr fields, 'author_name', 'timestamp', 'message', 'source_id', 'no_of_followers', 'tag_names' are now usable in Solr queries, for example, below is the search command of "GlocoMRIMachine" on all messages:

```
'http://gloco.com:8983/solr/glocosocialmedia/select?q=GlocoMRIMachine&wt=json&indent=true&rows=10
{ "responseHeader":{
   "status":0,
   "QTime":5,
   "params":{
    "q":"GlocoMRIMachine",
    "indent":"true",
    "rows":"10",
    "wt":"json"}},
 "response":{"numFound":102,"start":0,"docs":[
    {    "source_id":"tag:search.twitter.com,2005:628243496017768449",
      "message":"GlocoMRIMachine works like a charm",
      "author_name":"Saint Peters Hospital",
      "timestamp":"2016-03-15T12:18:25Z",
      "no_of_followers":690,
      "id":"278536b6-801e-4d7f-a22a-7d1558c6b00a",},
```

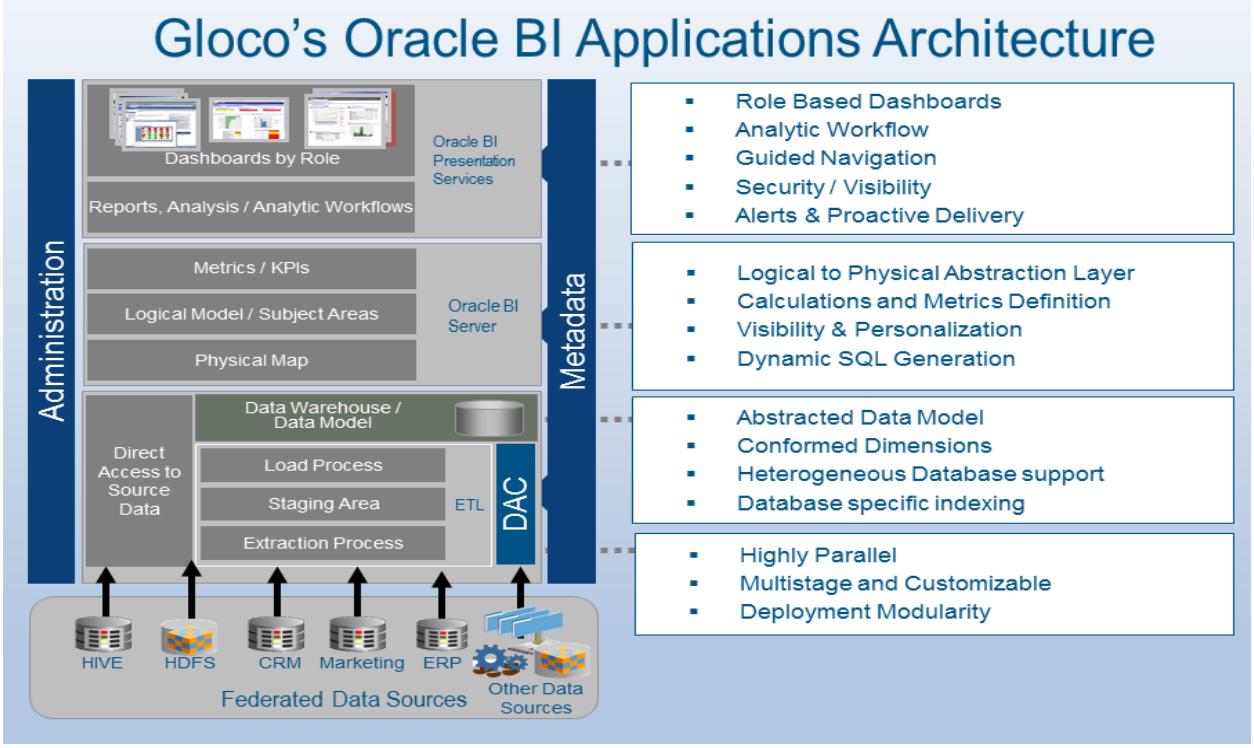
A snapshot of Solr's built-in user interface is given below:



The job will be configured in Oozie workflow to purge the indexed documents on a weekly basis.

3.5 INTEGRATION AND VISUALIZATION DETAILS

Data Integration Steps Using Data Warehouse



The Informatica data warehouse and Oracle Business Intelligence Enterprise Edition (OBIEE) analytics tools will be the primary infrastructure involved in the integration and subsequent visualization of the social data. These tools are a part of the Gloco on-premise BI suite. These products are currently used for visualization of ERP, CRM and marketing database reports at Gloco.

In order to connect the Hive external tables for the social media analytics solution with the Informatics data warehouse, the following steps will be followed:

- Establish connection with the Hive tables using Informatica connecter for Hadoop.
- Import the table metadata from the Hive table to the Informatica data warehouse and create data objects for sources and targets that need to be used in a mapping.
- Configure the source definition in source analyzer of Informatica designer client for imported data objects.

Integration Process with GLOCO CRM, ERP

In order to integrate the social media tables which were ported to the data warehouse with the enterprise data, which might reside in multiple disparate source systems such as relational databases like CRM or ERP or Marketing database, we need to follow these steps:

- Establish connection with the CRM tables using ODBC drivers
- Import the table metadata from the CRM accounts and contacts tables to the Informatica data warehouse. Also create data objects for sources and targets that need to be used in a mapping.
- Create a logical data model to define the relationship of the social media tables with the imported CRM objects.
- Create physical objects for storing the target outputs based on source qualifier.
- Configure the relationships between the imported logical data objects and physical data objects in mapping designer.

CRM matching criteria

Based on the following matching criteria we correlate between the tables:

- Transformation lookups will be done to match the customer name data to the CRM account name.
- Transformation lookups will be done to match the author name data to the CRM contact name
- Transformation lookups will be done to match the customer location data to the CRM address information.
- Based on the matching algorithm a matching score will be assigned to the record, which helps the user identify if the customer exists in the CRM database.
- Configure the transformations between the imported logical data objects, lookups and physical data objects like dimension and FACT tables.
- Create a session for each mapping in workflow designer tool and validate the workflow for the output.

Scheduling of Jobs

Once the workflow is ready the next step is to plan and schedule the execution using Informatica Dynamic Access Control (DAC).

- Create execution plan for scheduling job based on the intervals on every hourly data load of our social media Hive data.

- Jobs are separated for full load of data and incremental load of data, incremental data load will based change data capture(CDC) on source qualifier filter on new data set which adds only the deferential data has been loaded from our social media Hive data.
- DAC can automate these jobs based on regular intervals or in incremental
- You can also use the Monitoring tool to view logs for workflow instances and to view workflow reports.

At the end of this process the Gloco social media analytics solution will have FACT and dimension tables created and ready for visualization.

Visualization Process

The Gloco social media analytics solution will employ the OBIEE suite of BI presentation applications to display the analytics dashboards. Oracle BI Presentation Services provides a rich interactive user experience within a Web environment based on HTML, DHTML, and JavaScript. Gloco marketing will be able to get near real-time web analytics refresh on reports using BI. This will help the users access real-time content and quickly optimize depending on who is talking about the Gloco products. Clear and descriptive data visualization is essential for the interpretation of business data use cases to identify trends in social media or outliers in real-time and guide exploratory data analysis. Oracle BI simplifies the creation of visualizations reports and provides powerful visualization facilities.

OBIEE Dashboards

Web Catalog presentation: Oracle BI analytics presentation services provide a pure browser-based administration tool to administer all necessary functions in the Web Catalog. Oracle BI presentation services complement Oracle BI Server security with extensive set of controls, configuring privileges to access functionality in the Oracle business intelligence UI. BI administrators can control which users can access what dashboards and set user privileges; create and manage groups and BI access roles; re-name or delete saved analyses, and view and manage sessions. After saving analyses, users can easily add these complex layouts to Dashboards using a drag and drop dashboard editor to share these publicly. Dashboards can be tweaked and modified without limit. So for instance to display the heat map for the first use case the Gloco social analytics OBIEE dashboard can leverage fields such as sentiment, geo-location and company name to build the report. Additionally trend reports, conversations and influence reports can be generated using similar fields.

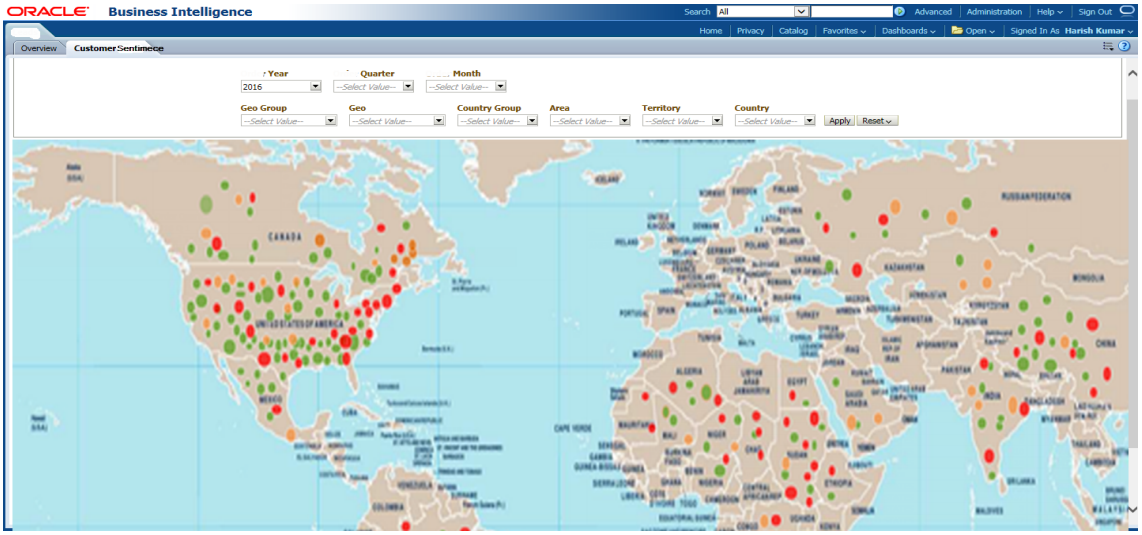
Connecting OBIEE with social analytics data

The Oracle BI interface generates logical SQL, based on report criteria chosen by the user, the results of which can then be formatted, displayed in the Oracle BI interface. Based on the data request from the BI server the OBIEE will pick the data from FACT and/or Dimension tables.

3.6 SOLUTION DEMONSTRATION

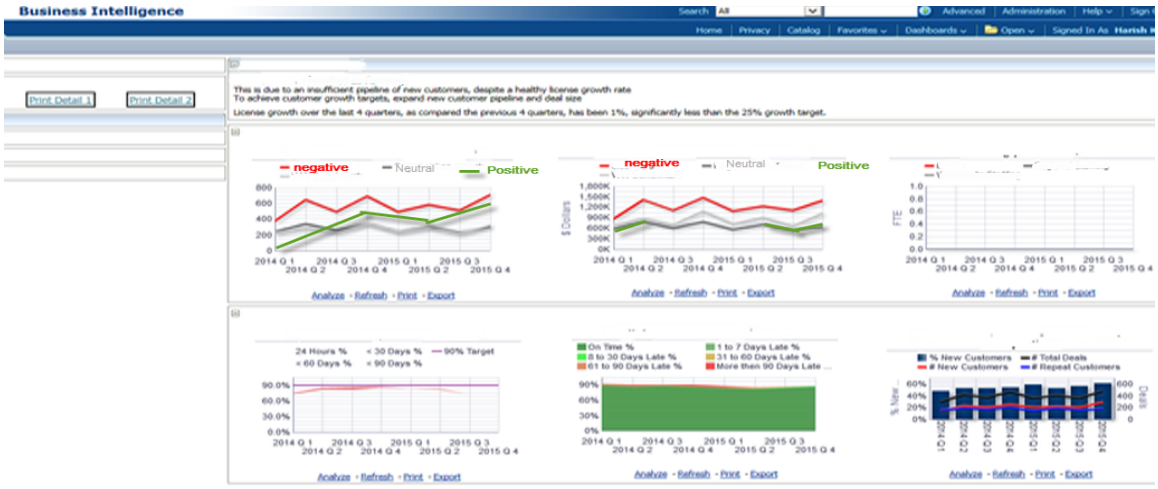
Use Case 1 Heat map

To generate a heat map for use case 1 the OBIEE application will use the BI geospatial views. BI Map views can be inserted into any analysis and are presented as multiple BI analytics data layers mapped. Map Views are generated by Oracle Map viewer, which is a Java application and map-rendering feature of BI Application Server.



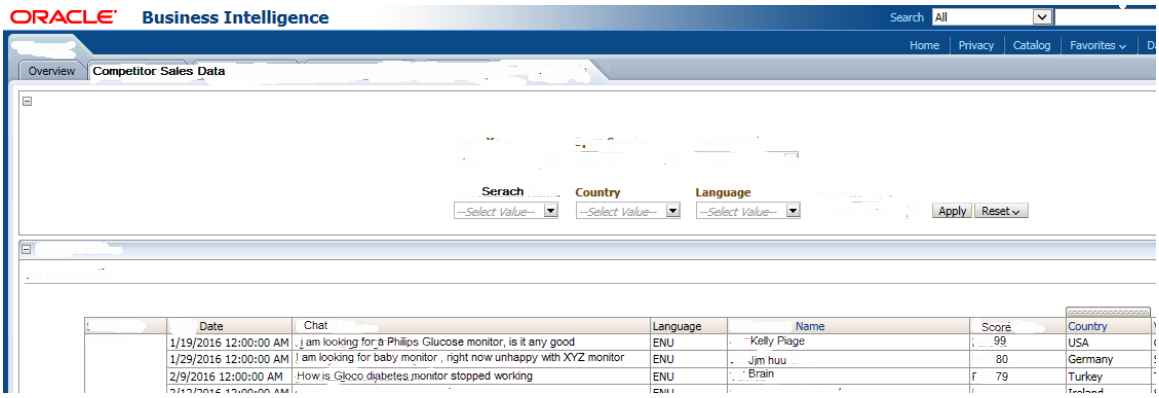
Use Case 2 Trend Report

To generate charts and graphs for use case 2, BI Analysis & Reporting allows dozens of out-of-the box graph views to be added and customized based on user requirement. Using pivot tables we can slice and dice data. Data visualization components are a set of rich interactive JSF components that provide animation, interactivity, graphical and tabular capabilities for visualizing/analyzing data. Use case 2 will display trend reports that track sentiment by time.



Use Case 3 Conversation analysis

Gloco provides predictive element to analytics users, which helps digital marketing anticipate the actions that will be taken based on customer social media conversations. To generate dashboards that display user conversations by filtering criteria, OBIEE combines filters and geospatial tools. The user can search conversations by Gloco product name or competitor company name with the added advantage of being able to tie it to customer profiles obtained from the CRM.



3.7 DEVELOPMENT PLATFORM

ICT Development team at Gloco requires the development process that provides better engagement with business users, shorter predicable delivery and is flexible to changes. An Agile based methodology will be used for managing development and controlling iterative and incremental software releases.

The following development tools will be used

- Eclipse IDE for developing Map Reduce programs using Hadoop libraries and Java 8.
- GNIP and DataSift client jars will be integrated into the application development environment.
- Jenkins will be used as a Continuous integration tool which is integrated with the GitHub for source code repository, maven for builds, Nexus repository for third party software, JUnit for regression tests, Cobertura for code coverage, Sonar for code analysis.
- Jenkins will be configured to push the deployable artifacts to test, integration and production environments that are hosted on the cloud servers.
- Virtual Machines in the cloud are set up with Cloudera CDH distribution software 5.0. that includes all the required software for this project, such as Apache Flume, Apache Hadoop, Cloudera Search powered by Solr, Apache Hive and Java 8.
- The software increment is pushed to production once the user acceptance test is completed and accepted by the stake holders and product owner.
- Existing Informatica data warehouse will be used for developing the integration solutions with Social media data and CRM/ERP/Marketing.
- Existing OBIEE suite will be used for developing the visualization of reports.

3.8 SYSTEM METRICS

#	Metric	Value

1	No of business users	200
2	No of administrative users	5
3	Yearly data volume	1 to 2 TB
4	Dash board response time	Less than 30 sec
5	System availability	24/7
6	Sentiment classification accuracy	Accuracy rate of 75%. Measured using random samplings

3.9 VENDOR SELECTION CRITERIA

Software	Comparison	Description	Selection
Social media data service providers	Gnip	Only service provider vendor for Twitter firehose social media enterprise data.	Gnip
	Datasift	Datasift is the only provider for Facebook topic data.	Datasift
Cloud based Software	Amazon	Amazon provides the cost effective option for storage for Hadoop cluster of servers, offers more community support and market share.	Amazon
	Azure	Compare to Azure, Amazon is more mature and has more options of pricing on storage.	
	Premises Cloud	This is the most expensive option, which may not provide on-demand scalability.	
Hadoop stack	Cloudera	CDH has more options, is enterprise scalable, more mature since available in the market from 2008, and offers more community support.	CDH
	HortonWorks	Hortonworks is relatively new, and in market since 2011 with less community support and lower market share.	
Search	Solr search	Powered by CDH, and very fast compare to other search options	Solr
Integration	Informatica	Existing Informatica is cost effective when compare to the CDH stack software like Cloudera Impala. In-house expertise is available for building the solution. Data warehouse servers are readily available	Informatica
	Impala	There is additional license fee for the installation and usage, in-house expertise is not available	
Visualization	OBIEE	Existing OBIEE is cost effective when compare to the CDH Spark. In house expertise is available for building the solution. BI Servers are readily available.	OBIEE

Architectural Principles Revisited

- The architecture is flexible enough to accommodate any number of firehose-based sources of data provided by Gnip and DataSift. This architecture document illustrates Twitter and Facebook as the primary data sources, since they are the leading social networks around.
- The architecture aggregates, structures and formats the data using technologies such as Flume, HDFS, HCatalog, Hive and Oozie.
- The solution supports integration with existing Gloco CRM systems.
- The data is modeled into analytic objects, fact and dimension tables for visualization using OBIEE suite of BI presentation applications.
- The dashboards present intuitive and actionable business intelligence, which will help Gloco achieve its business goals.

4 IMPLEMENTATION PLAN

The implementation plan describes the development, deployment, and quantification of our solution, and the success metrics used to validate it. The following sections are covered:

- Solution development describes the development approach, timeline, and milestones.
- Solution deployment describes the deployment approach and the implementation resources.
- The Operationalization section describes the supporting non-functional processes required to run the system along with the user enablement process, which includes devising a training structure and user workflows to ready the users for the system.
- The Success metrics explain how the success results will be measured and reported.

The Gloco ICT will be responsible for the implementation and delivery of the solution.

4.1 SOLUTION DEVELOPMENT

Development and deployment of the analytics portal and related project tasks will be completed in a phased approach. The Gloco ICT implementation team and a cross-functional team will facilitate the meetings, communications, and adherence to the project plan. Weekly status reports will be created by the Project Management Office (PMO) to inform stakeholders on progress of the project.

The implementation for the entire analytical system is a nine-month program with four phases of implementation.

Phase	Milestone
Phase 1	Data sources integration for Twitter and Facebook social media data is complete. Infrastructure set up is complete in lower environments. Business users will be able to configure the filter rules.

Phase 2	Administration support to view and search raw social media data will be available for administrator and Data Specialist users. Work Flow set up and configuration for the jobs that process social media data is complete.
Phase 3	Business users will be able to visually track customer sentiments by time using Trend Reports. Business users will be able to oversee customer sentiment across the world using Heat Map Dashboards. User Training Session 1 is complete
Phase 4	Business users will be able to track conversations and find top influencers of Gloco and its competitors via the Dashboard for Conversation analysis. Integration with Internal Enterprise applications is complete. User Training Session2 is complete Solution goes live for operational use.

ICT Assumptions

- It is assumed that the project is initiated and the recommended hardware is purchased before starting first sprint.
- It is assumed that ICT group is trained and skilled with DevOps methodology using Gloco's standard development tools such as Continuous integration Jenkins, Selenium, sonar cube. New developers and testers will be trained with DevOps principles using webinars.

Development Strategy

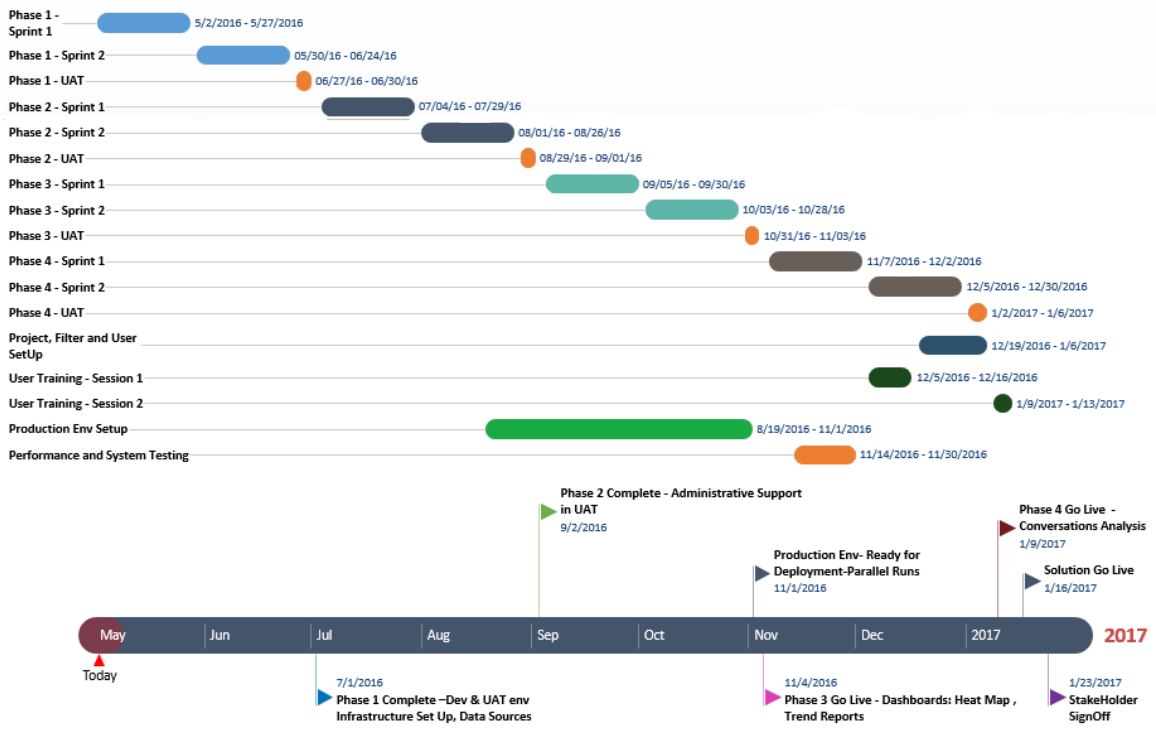
An Agile methodology has been chosen to design, develop, and perform continuous integration and delivery. Multiple sprints will be developed and delivered in each phase and deployed as an increment to UAT. The project time line given below shows both development and delivery schedule along with the supporting activities. For continuous planning, development and integration we will utilize DevOps principles, however in the interest of saving infrastructure costs, the deployment to production occurs once phase 3 is complete. Each sprint will start with sprint planning to define the features as backlog items, which will be managed and tracked in the tool Jira. The development team will comprise of three sub-teams with a scrum master, lead developer, three senior developers, a business analyst and a tester. List of backlog items will be distributed among the team members of each sub-team to deliver the output in a fixed time frame of 4 weeks. After each sprint is completed, a sprint demo will be presented to the stakeholders and the product owner. Based on the feedback from the demo, new backlog items may be added and prioritized for the next sprint. Stakeholders and product owners will provide sign off on the increment.

The following table summarizes the expected delivery of each sprint from both technology development and supporting services.

Phase/Sprints	Development Deliverables	Supporting Deliverables
Phase 1-Sprint 1: Infrastructure set up and Development Env set up.	1. Set up Continuous integration tool Jenkins and integrate with Sonar Cube, Maven, JUnit, GitHub, Corbetura for Code coverage etc. 2. Build common core framework for Java Map Reduce jobs and Hadoop libraries to help developers build upon using eclipse IDE. 3. Install Cloudera stack on DEV VM.	• Deployment diagram: IT hardware, software, and networking components • Development platform is complete. • Physical deployment of hardware on DEV VM environment is completed and required software is installed.
Phase 1 -Sprint 2: Data Source Integration and Rules administration UI.	1. Develop an interface to accept a configured set of rules and filters to be pushed to Data Sift and Gnip to filter and classify Gloco's data. 2. Develop an interface to stream social media data from Data Sift and Gnip Data sources and store the messages on HDFS. 3. Develop a user interface for Data specialist to configure rules and filter criteria.	• Filters and Rules are configured externally to the code and documented. • Rules administration screens are available for users to configure rules and filter criteria.
Phase 1 will be deployed to UAT environment for functional and UAT testing. Stakeholders provide sign-off on Phase 1. Infrastructure set up and connectivity is validated in lower environments.		
PHASE 1 is COMPLETE		
Phase 2-Sprint 1: Map Reduce Jobs	1. Configure Oozie Work flow and build scripts to automate the configuration. 2. Develop and build Java Map reduce job to perform dictionary set up, Solr search indexing. 3. Develop and build Hive Script jobs	• End to end Oozie workflow is complete and automated. • User is able to search and view raw social media data.
Phase 2-Sprint 2: Map Reduce Jobs	1. Develop and build Java Map reduce job to perform Sentimental analysis, data clean up jobs etc. 2. Develop and build Hive Script jobs	• Data mining process of social media data is complete on HDFS.
Phase 2 will be deployed to UAT environment for functional and UAT testing. Stakeholders provide sign-off.		
PHASE 2 is COMPLETE		
Phase 3 - Sprint 1: Data Warehouse Integration	1. Integrate Hive database to Data Warehouse using ETL processes.	• Structured Social media data persisted in Data Warehouse
Phase 3 -Sprint 2: Dashboards-Heat Map and Trend Reports	1. Build algorithm to aggregate data required for trend reports. 2. Build algorithm to aggregate data required by Geo location for heat map.	• Dash board view of Trend reports and Heat Map completed
Phase 3 will be deployed to UAT environment for functional and UAT testing. Stakeholders provide sign-off and increment is deployed to Production.		
PHASE 3 is COMPLETE → GO LIVE to Production		

Phase 4 -Sprint 1: CRM integration	1. Integrate with CRM 2. Integrate with Supply chain Management 3. Build and develop a job to update the score based on the customer matching criteria.	• Gloco CRM Customers and social media users are linked together with an assigned matching score.
• Due to the parallel runs in phase 3 the defects that are found during testing on production will be scheduled into an additional sprint and will be deployed to production.		
Phase 4 -Sprint 2: Dashboard for Conversation Analysis	1. Build dashboard required for Conversation analysis Use case 3	• Users can view and search conversations and relate to existing customers of Gloco.
Phase 4 will be deployed to UAT environment for functional and UAT testing. Stakeholders provide sign-off and increment is deployed to Production.		
PHASE 4 is COMPLETE → GO LIVE to Production		

Timeline:



4.2 SOLUTION DEPLOYMENT

Production environment setup tasks will start after Phase 2 of the project is completed and production environment will be ready for code deployment by Phase 3. Since the new solution is an automated system replacing a manual system, Gloco ICT will use the parallel conversion approach to deploy the solution. The current AS-IS manual process will continue until the project sponsor signs off the new solution. Parallel runs will start with Phase 3 of

the project. Multiple teams will be formulated from Gloco ICT to handle the solution deployment tasks. Coordination amongst the teams and communication with stakeholders and business users are very important for solution go-live. The resources needed for production environment setup and solution deployment are listed below.

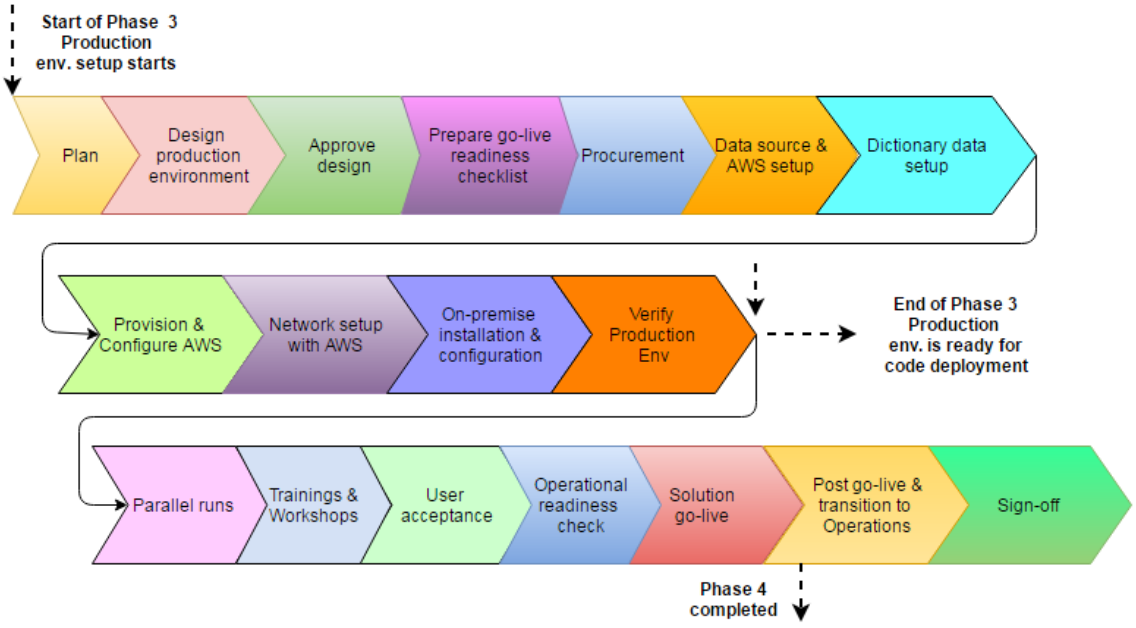
#	Project Role	Team Name	ICT Role	Responsibility
1	Project Sponsor	PMO	VP of Marketing	Provide executive support and secure project resources and approvals. Final sign-off
2	Project Manager	PMO	VP of ICT PMO	Overall manager of the project. Coordinates with ICT and Business departments
3	Business Analyst	Data	ICT System Analyst	Requirements management and communication
4	Enterprise Architect	Architecture	ICT Enterprise Architect	Ensure the solution fits with ICT technical architecture.
5	Solution Architect	Architecture	ICT Application Architect	Overall design of the solution
6	Service Delivery Manager	Systems	ICT Operation	Manages infrastructure build and solution deployment
7	System Administrator	Operations	ICT systems team member	Manages System support
8	Network Administrator	Operations	ICT systems team member	Manages Network support
9	Database Administrator	Operations	ICT systems team member	Manages database support
10	Technical Support	Operations	ICT systems team member	Provides system support
11	AWS System Administrator	Systems	ICT systems team member	Manages AWS system administration
12	Security Administrator	Security	ICT information security team member	Ensures ICT security standards are met.
13	Configuration Management Administrator	Systems	ICT Operation - CM Admin	Ensures that changes are properly approved and communicated
14	Lead Developer	Development	ICT Application Delivery	Scrum master as well as the lead for sub-team
15	Senior developer	Development	ICT Application Delivery	Sub-team member, developer
16	Functional Tester	Testing	ICT Application Delivery	Sub-team member, embedded tester
17	Performance Tester	Testing	ICT Systems	Test solution meets non-functional requirements

Following table shows the production environment setup solution deployment tasks and the primary team responsible for the task.

#	Task	Resource	Duration (days)	Deliverables
1	Plan deployment process	PMO team	5	Overall go-live plan document.
2	Design production environment	Architecture team	15	Detailed deployment architecture document.

3	Prepare production readiness checklist	PMO team	5	An overall checklist document.
4	Procurement for on premise infrastructure hardware/software	PMO team	20	List of hardware/software. Purchase Order issued.
5	Data Source Setup	Systems team	5	Signed contracts documents with Gnip and Datasift. Configuration details documentation.
6	AWS Account Setup	AWS team	5	Signed AWS Customer agreement. Configuration details documentation.
7	Dictionary data setup	Data team	15	Corpus for sentiment analysis ready.
8	Provision and configure AWS CDH instances	AWS team	5	Configuration document. CDH instances ready. Configuration details documentation.
9	Network setup between AWS and Gloco	Systems team	5	Configuration document. Working network between AWS and Gloco. Configuration details documentation.
10	Install & Configure on premise infrastructure for data warehouse and OBIEE	Systems team	5	Configuration details documentation. Infrastructure for Integration and visualization layers ready.
11	Verify Production environment	Systems team	5	Use the checklist to verify production environment is ready for code deployment.
12	Build non-functional requirements test plan.	Testing team	10	Performance/System test plan document.
13	Provision end users	Systems team	5	End user accounts created.
14	Parallel runs	All teams	30	Solution deployed. Communication document and plan.
15	Sysadmin/operations trainings	Systems team AWS team	5	ICT Operations trained.
16	User trainings and workshops	User training team	15	All business users trained.
17	User acceptance	Business team	5	User acceptance document signed.
18	Operational readiness check	All teams	2	Verified checklist.
19	Go-live	Systems, Operations, Business and PMO teams	2	Complete solution deployed and working.
20	Post go-live & transition to operations	Systems & Operations team	10	Stable system.
21	Sign-off	PMO	1	Project sign-off document.

Production deployment tasks:

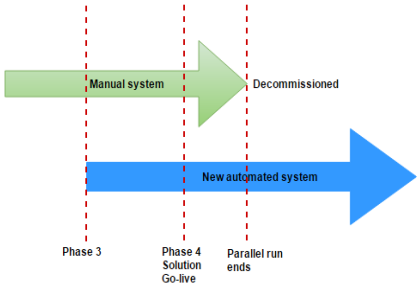


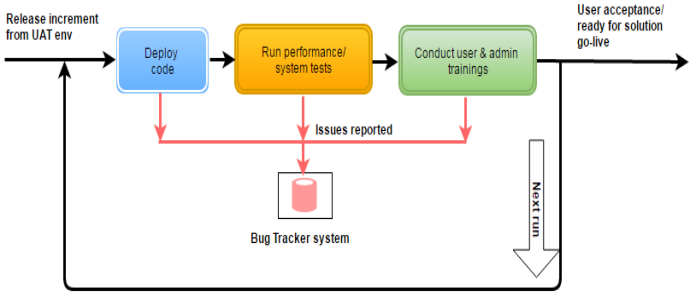
All tasks except parallel run are one-time tasks to setup the production environment.

Parallel runs with current AS-IS process.

After the production environment is ready, increment code deployments on production will start with end of Phase 3. There will be multiple parallel runs of the solution where each increment from UAT environment will be deployed to production. Deployment to production will occur at the end of each phase. Performance and systems testing will be done using a cross functional team from ICT and marketing department. ICT will use a bug tracking system to report both functional and technical issues to development team. After final user acceptance, solution will be deployed for go-live in Phase 4 of the project. After solution go-live, team members will help resolve issues until the solution is operationalized. After this period, the solution will be transitioned to ICT operations for ongoing support.

Following diagrams show parallel run and the processes within parallel run.





4.3 OPERATIONALIZATION OF THE SOLUTION

The social media solution will be systematically integrated into the Gloco operational platform. The embedding of the social media analytics solution will integrate actionable social media insights into the existing Gloco decision-making systems and business processes. As a part of operationalizing of the solution, we identify the supporting non-functional components that will help run the social analytics solution.

Supporting Non-Functional Processes

The following non-functional components will help operate the social analytics solution:

User Administration Console
A user administration console will be created to setup user roles and permissions as well as configure the social media analytics solution. Once signed in, the users can create a new project and for each project, access the configuration details for filtering, reporting and analysis.

Security and Access Management Process
Facilitation and control of user access to the social analytics solution will be accomplished by integration with the Gloco corporate portal and enablement of Single Sign-On (SSO) services. The social analytics solution will be integrated with Gloco's standard identity and access management service facilitated by Oracle Access Management (OAM 11g). The OAM will check the user credentials against Active Directory and allow the user to navigate to the social analytics solution via the corporate portal. Different user groups will gain access to the portal by following a standard Service Request (SR) process established at Gloco. The process of setting up the SSO for a user role is detailed below:

- The user needing access to the social analytics portal will submit a SR, requesting access to one of the pre-defined roles.
- Once the SR is approved, it gets routed to the Super Administrator who will use the user administration screens to locate the user and assign the right policy in the Active Directory.
- The administrator will then setup user roles such as Customer Strategist, Data Specialist etc. and their corresponding permissions.

SSO will thus be enabled for the users and the analytics solution and will be accessible via the corporate portal.

HDFS Administration
The Cloudera manager server will host the admin console that will help manage the CDH cluster. The Data Specialist role will be responsible for using the Cloudera manager console to install, configure, start and stop the services needed to operationalize the cluster. Solr search can further aid in the administration of the raw data stored on HDFS.

Data Sources Administration
A Data Fetcher Development Kit will provide the ability to integrate data sources outside of the DataSift and Gnip firehose sources. The kit will define specifications on how to code and

test custom data fetcher classes that can interact with an external source, fetch and convert the data into JSON format, and have it be consumed by the Gloco social analytics solution.

System Performance Monitoring

- The Keynote testing and monitoring tool will simulate synthetic transactions to generate dashboard views, thus mimicking an end user and periodically ensuring the system is up and running.
- Performance of the dashboards will be measured against a threshold (5 mins.) and any degradation will be notified to the support team. The system should be able to handle 100 concurrent simulated users with a 30 seconds response time.
- For the AWS instances, alarms setup in CloudWatch can track metrics and auto scale instances based on the load.

Change Management

The change control process at Gloco establishes how the changes will be proposed, accepted, and controlled. As a specific example, advances in the sentiment analysis techniques might result in the need to incorporate an enhanced machine-learning algorithm into the analysis workflow. The change control process in this instance would adhere to the following steps:

- Classify machine-learning algorithms by regression, classification, clustering, and anomaly detection. Identify, categorize and choose the appropriate algorithm.
- Evaluate the chosen algorithm on factors such as accuracy, training time, linearity, number of parameters, number of features and special case detection capability.
- Evaluate all the relevant metrics. Generate a report that details the impact and results of the evaluation.
- Submit the report and change request to the Gloco ICT change authorization board (CAB) and obtain approval. Schedule the change request for the upcoming release and follow the development cycle.
- Incorporate the algorithm and corresponding changes into production.

4.4 USER ENABLEMENT

The following criteria will be used to ensure user readiness:

- Training is readily available for all aspects of the social analytics solution.
- Users are made aware of the changes and updates to the social analytics solution.
- Participation and feedback from the pilot usage and training sessions are well documented.
- All issues and questions are addressed over the course of the user readiness activities.

As the system enters production, the internal user roles will be trained and familiarized with roles, responsibilities and operational specifics of the solution. A pilot solution will be set up for training purposes.

Training

User Group	Training Scope	Training Method
Administrators: Super- Administrator Data Specialists	User creation, project creation, configuration, administration of processes and tools from data ingestion to visualization.	Documentation, user-guides, how-to-videos, meetings and conferences.
Business Users: Marketing strategist, Customer strategist, Product strategist, Sales strategist.	Relevant business user workflows according to role.	Solution Demos, Instructor webinars, pilot evaluation groups, classroom training.
Installation Support: Infrastructure Engineer	Setup and maintenance related to Flume, Hadoop, Hive, MapReduce, OBIEE servers, OBIEE dashboards etc.	Documentation, user-guides, how-to-videos, meetings and conferences.
Operational Support: Operational Support Team	Solution operations process including monitoring tools.	Documentation, user-guides, how-to-videos, meetings and conferences.

4.5 PROJECT RISKS

Risks	Impact	Probability	Mitigation Strategy
Project is solely dependent on third party feeds from Gnip and Datasift	High	Low	Explore contingency options for the downtime including other data providers.
Change in the data feed regulations by providers like Facebook or Twitter	High	High	Changes in the data provider regulations would directly impact the business. Explore additional social media data sources.
Cloud dependency for processing large input files from third party	High	Medium	Make sure the cloud infrastructure is elastic and geographically replicated.
ETL connectivity from cloud to in-house network connectivity	High	High	Largely depends on the ISP Vendors, network failures or disruptions. ISP's should notify the teams. The teams should be responsible for the communication with the internal analytics users who access the reports.
Cost of maintaining AWS infrastructure can go unexpectedly high.	Medium	High	Ensure alarms are setup to monitor AWS usage and the proper data clean up mechanism is performed at regular intervals.

4.6 SUCCESS METRICS

GLOCO will measure the success of the implemented solution by using four Key Performance Indicators derived based on the business goals. A total of 10 metrics scorecard will be tracked and reported using Analytics reporting. The predictive analytics solution

will have the ability to populate these metrics on a daily basis using the data retrieved from social media analytics and CRM systems (Refer: Reporting Scorecard). The metrics will be reported at a company level and can be drilled down to the detailed level for greater granularity. The KPI metrics can be tracked from the marketing scorecard and the business KPI metrics can be viewed in dashboards.

KPI Category	Success metrics	Target (Score card)	Metric Collection	Time-line
Improve customer experience	Based on the number of positive comments, improve the customer experience by 5%	15% ↑ (refer Scorecard)	By using Analytics tools provided to track and engage social media conversations with customers, there will be an increase in 15% change in customer satisfaction.	1 year
Improve customer service experience with the product	10% reduction in customer complaints. 5% improvement in customer services.	10% ↑ (refer Scorecard)	Based on tracking and identifying the social media conversations about Gloco product defects and complaints the customer strategist will engage with the user, resulting in a 10% revenue increase.	1 year
Increasing customer proclivity to buy a particular product by region or demographic.	Increase in product sales by 5%	5% ↑ (refer Scorecard)	Based on the product sentiment trend in a particular region we will engage with the product performance issues resulting in increased sales.	1 year
Improve interaction with customers and identify ways to engage directly with them.	5% increase in lead generation and 2% increase in opportunity lead to sales closeouts. 3% increase in prospects and 1% in revenue.	20% ↑ (refer Scorecard)	Customer engagement has helped marketing users find leads resulting in new product sales increasing sales by 20%. Product strategist have incorporated new ideas and added to the product line, which has resulted in 5% increase in revenue.	1 year
Raising Brand health and reputation, social media avenues and marketing channels to reach out to the consumer.	5% Increase in brand satisfaction. 2% reduction in marketing expenditures in campaign outreach because of improved health of brand.	10% ↑ (refer Scorecard)	Brand health and trend analysis dashboards provide heat map and sentiment analysis to represent the data based on the geo-location.	1 year

Scorecard Measurement

#	Perspective	Measurements

1	Financial	Return of Investment, Economic value added by project, new sales revenues
2	Customer	Customer health, Brand equity increased, market share difference, Positive aspects of company name
3	Business Process	• Measuring the improvement from manual AS-IS process to automated social media conversation monitoring. • Negative aspect comments that have been followed up. Engagement from marketing team has made customer satisfactory/happy. • Product defect reported by customer that has been followed up and closed. • Social Media Conversations tracking that has created leads and prospects, which generated marketing and sales opportunities.
4	Learning	• Training for the internal users like sales strategist, product strategist, marketing strategist • More refine way to deal with customers, which is part of continuous improvement

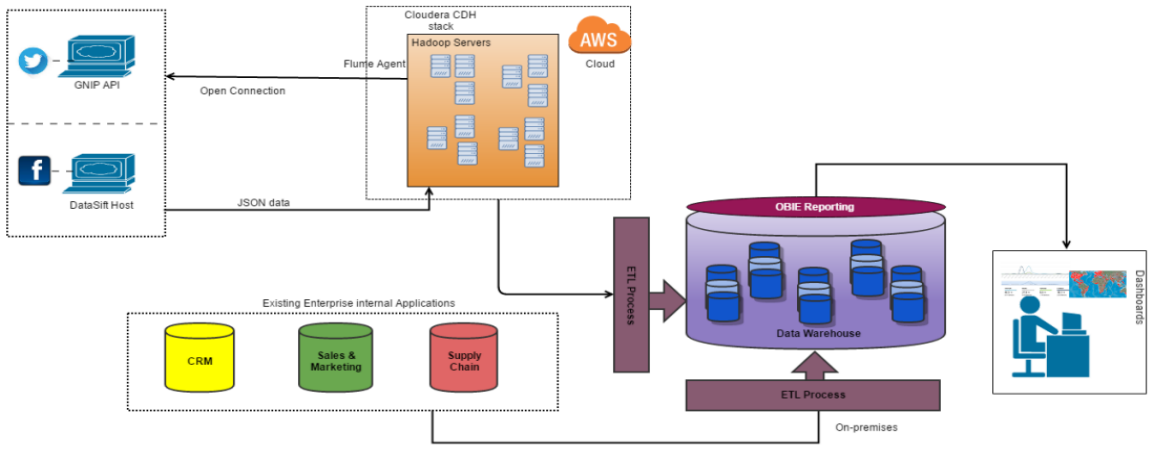
Metrics Collection Scorecard

Social Media Metrics Reports	Manual Process before implementation	After Go-Live Year 1	After Go-Live Year 2	% Change
Number of followers in social media	15000	50000	65000	15
Number of people re-tweeted our company messages	3000	5000	6000	10
Number of people liked our company, products	5000	10000	12000	25
Number of times brand company name was mentioned in positive aspect	0	5000	5500	10
Number of times brand company name was mentioned in negative aspect	0	3000	2700	(10)
Number of product defect reported being followed up by Service team	20	400	440	10
How many times company products were mentioned on positive aspect	0	8000	8800	10
How many times company products were mentioned in negative aspect	0	600	540	(10)
How many times marketing agency followed up the brand negative aspect	0	500	600	20
Number of prospects, leads generated by monitoring conversations	0	600	720	20
Number of leads generated according to trend analysis for particular products by performance in regions.	0	400	420	5

5 APPENDICES

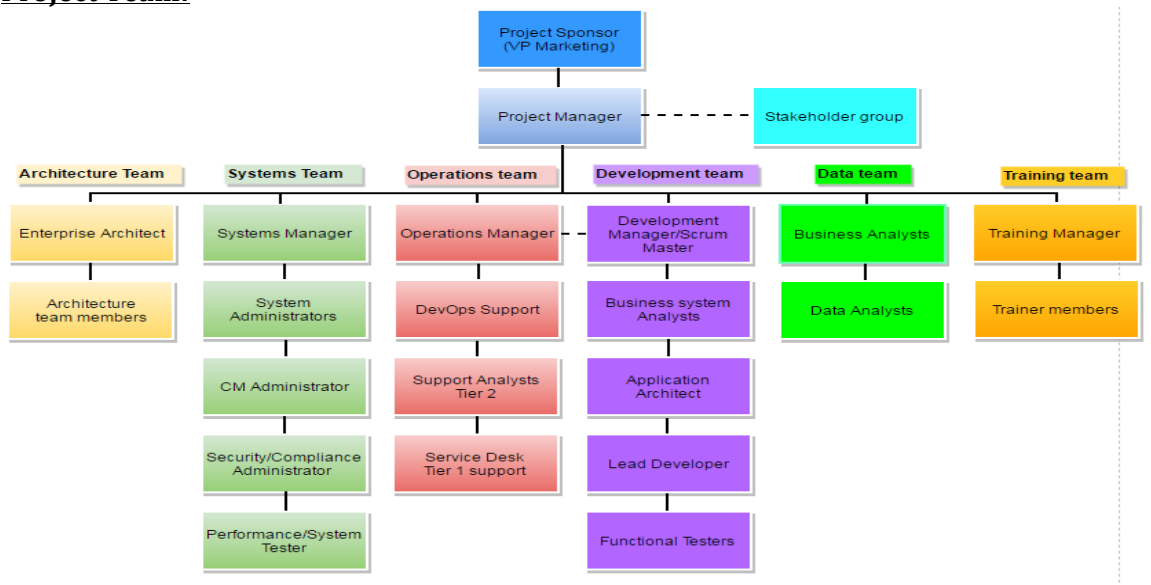
APPENDIX: A

Deployment Diagram:



APPENDIX: B

Project Team:



6 REFERENCES

1. Sussin, Jenny. (2015). Market Guide: Social Analytics applications for IT leaders. Retrieved from Gartner. Article: G00279553.

2. Natkins, John. (2012). How to analyze Twitter Data with Apache Hadoop. Retrieved from https://blog.cloudera.com/blog/2012/09/analyzing-twitter-data-with-hadoop/
3. Ford, K. (2015, November 18). Facebook Topic Data Just Got an Upgrade with Super Public Text Samples. Retrieved from http://blog.datasift.com/2015/11/18/facebook-topic-data-just-got-an-upgrade-with-super-public-text-samples/
4. PowerTrack Rules. (n.d.). Retrieved from http://support.gnip.com/apis/powertrack/rules.html
5. PowerTrack Rules. (n.d.). Retrieved from http://support.gnip.com/apis/powertrack/rules.html
6. Data Format. (n.d.). Retrieved from http://support.gnip.com/sources/twitter/data_format.html#SamplePayloads
7. Why Hadoop? Analyzing Social Media Sentiment Data. (n.d.). Retrieved from http://hortonworks.com/use-cases/sentiment-analysis-hadoop-example/
8. PYLON for Facebook Topic Data | DataSift. (n.d.). Retrieved from http://datasift.com/products/pylon-for-facebook-topic-data/
9. Solr Features. (n.d.). Retrieved from http://lucene.apache.org/solr/features.html
10. Cloudera Search. (n.d.). Retrieved from http://www.cloudera.com/documentation/archive/search/1-3-0/Cloudera-Search-Installation-Guide/csig_install_search.html
11. Informatica Cloud Hadoop Connector Guide. (n.d.). Retrieved from https://network.informatica.com/docs/DOC-15580