Chapter 12
Data Analysis for Potential Reach Using Elasticsearch

Nikhil Chaudhari
Vellore Institute of Technology (VIT), India

ABSTRACT

Elasticsearch has become an attractive open-source search and analytics engine for use cases such as log analytics, click stream analytics and real-time application monitoring. As a service, Elasticsearch is made available by Amazon Web services, Searchly.com, Bonsai and many other websites as a hosted engine. This hosted Elasticsearch service is known as Elastic Cloud. The aim of AWS (Amazon Web Services) is to provide Elasticsearch as a service to users. These web services can implement Service-oriented Architecture. In the following chapter Potential Reach is aggregated using the server. Potential Reach is a metric that indicates the reach of an online activity like tweet or comment. This number helps media marketers track the success of the brand or company.

INTRODUCTION

The potential reach metric allows to quantify not only the users, the company is engaged with, but also the followers of those users who may have seen company’s or a single user’s @handle or tweet. This chapter explains the data analysis method used by Elasticsearch to determine the Potential Reach of tweet. To begin we need to understand some of the esoteric terms.

Potential Reach

Potential reach is an important metric which indicates the spread of tweet or any other post or comment on social networking websites. Social networking websites have more use than just to entertain people in their free time. A new medium of communication has been opened to market analyst, marketing managers and publicity heads. Twitter is a green field for content marketers and social media managers. As of March 2016, Twitter has 310 million active users according to twitter growth statistics. Advertising has become important in today’s time to stay in competition. Many noted companies like Dell, Ford,
RackSpace use Twitter for publicity and small businesses also benefit from such websites as they can compete equally with large companies on social networking platform.

According to Simply measured complete guide to twitter analytics, potential reach is very useful number for social media marketer. This number is important because a key focus of social marketing is to expand audience and promote message to a wider segment of the population. The reach metric tells which content is working to grow audience and ultimately “reach” new people.

Potential reach for twitter, mathematically, is the number of followers of a user who tweeted, combined with the followers of users who re-tweeted the tweet. To understand better, here is an example: Suppose User “M” on twitter tweets about his brand using his own twitter handle. Now, the potential reach is equal to the number of followers of user “M” plus one(user “M”, himself). Subsequently, user “K” re-tweets user “M”’s tweet. This proliferates potential reach by adding the number of followers and one (user “K”) to the earlier total.

This chapter discusses how this metric was determined using data analysis on twitter sample data.

**Elasticsearch**

Elasticsearch is a search server. It can be used to search all kinds of documents. It provides scalable search, has near real-time search, and supports multi-tenancy. For these reasons Elasticsearch was chosen to determine potential reach. Elasticsearch uses Lucene and tries to make all its features available through the JSON and Java API.

Elasticsearch provides power of analytics with speed of search, which changes the relationship with data. The information gained can be used to improve products or change strategies accordingly. Also, Elasticsearch provides high availability and enormous growth according to amount of data.

Solr search platform was compared to Elasticsearch. It was found that Solr has very slow re-indexing and batch replication. Contrasting to this Elasticsearch has good API and is scalable according to need to application. This make Elasticsearch a favorite among users. According to Shay Banon, developer of Elasticsearch, the reason for its popularity is ability to communicate empathy to its users. In 2012, Elasticsearch BV was founded to provide commercial services and products around Elasticsearch and related software. And subsequently, these services got hosted online and came to be known as Elastic Cloud.

Recently, the Elastic Cloud has been made available to the consumers as a service. According to blog by Banon, Welcome Found, the company Found have created an extremely easy to use service with a strong technical foundation for Elasticsearch. This provides great service at affordable price and easy upgrading and scaling, with advanced security and useful plug-ins. Also, users get to choose SLA-based support from the creators of Elasticsearch.

**Kibana**

Kibana is open source data visualization plug-in for Elasticsearch. It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster. This plug-in is used to plot pie charts of information. Further mathematical transformations, slicing or dicing can be done on the information obtained.

Introduction to Kibana mentions that Kibana is great for real time data analytics and it allows users to search Elasticsearch data via Lucene Query Search String syntax. Lucene provides a rich query language through the Query Parser, a lexer which interprets a string into a Lucene Query using JavaCC. Lucene also supports Wildcard searches, Proximity searches, Range searches and Fuzzy searches.