2013 IEEE 16th International Conference on Computational Science and Engineering

5Ws Model for BigData Analysis and Visualization

Jinson Zhang
IEEE Member
School of Software, Faculty of Engineering & IT
University of Technology, Sydney
Sydney, Australia
Jinson.Zhang@uts.edu.au

Mao Lin Huang
School of Computer Software
Tianjin University, Tianjin, China
School of Software, Faculty of Engineering & IT
University of Technology, Sydney
Sydney, Australia
Mao.Huang@uts.edu.au

Abstract **— BigData, which contains image, video, text, audio and other forms of data, collected from multiple datasets, is difficult to process using traditional database management tools or applications. In this paper, we establish the 5Ws model by using 5Ws data dimension for BigData analysis and visualization. 5Ws data dimension stands for; What the data content is, Why the data occurred, Where the data came from, When the data occurred, Who received the data and How the data was transferred. This framework not only classifies BigData attributes and patterns, but also establishes density patterns that provide more analytical features. We use visual clustering to display data sending and receiving densities which demonstrate BigData patterns. The model is tested by using the network security ISCX2012 dataset. The experiment shows that this new model with clustered visualization can be efficiently used for BigData analysis and visualization.**

***Keywords** – BigData analysis; BigData pattern; data dimensions; data density; BigData visualization*

I. INTRODUCTION

BigData, according to Wikipedia (Aug 2013), is "the term for a collection of data set so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications". The datasets not only contain structured databases, but also include unstructured databases such as social media data or GPS (Global Positioning System) data.

According to Gartner 3Vs definition [2], BigData has three characteristics: volume, velocity and variety, as shown in Fig 1.

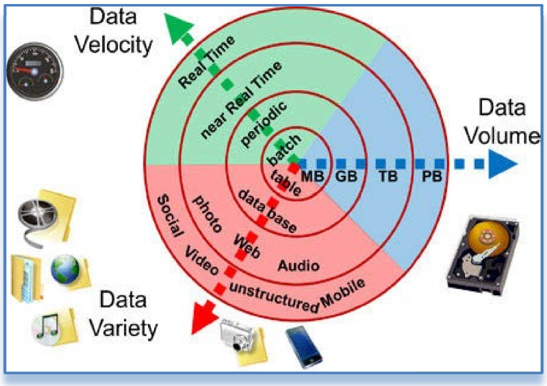


Figure 1. 3Vs model for BigData [3]

Volume describes datasets that are extremely large and easily amassed into Terabytes, even Zettabytes of information. Table I shown dataset volume size. Too much volume of dataset is not only a storage issue, but also a massive analysis issue.

TABLE I. DATASET VOLUME SIZE

VALUE	ABBREVIATION	NAME
1000^1	KB	Kilobytes
1000^2	MB	Megabytes
1000^3	GB	Gigabytes
1000^4	TB	Terabytes
1000^5	PB	Petabytes
1000^6	EB	Exabytes
1000^7	ZB	Zettabytes
1000^8	YB	Yottabytes

Velocity is how fast the dataset is being produced. Based on statistics from Pingdom 2012 [1], there are more than 38000 Google searches every second, 5 billion mobile phone users using 1.3 Exabytes of global mobile data traffic per month, 2.2 billion email users sending 144 billion emails per day, 2.7 billion likes on Facebook every day, 7 petabytes of photo content added on Facebook every month and 4 billion hours of video watched on YouTube a month.

Variety is how datasets contain both structured and unstructured data, such as documents, emails, audio files, images, videos, click streams, log files, or financial transactions. Hundreds, even thousands, of different attributes in multiple dimensions in the dataset provide too much information for traditional database management tools or applications to handle.

BigData comes from everywhere influence our life, and so is too big, too complex and moves too fast. For example, posting pictures and writing comments on Facebook; uploading and watching videos on YouTube; sending and receiving messages through smart phones; sending voice messages through WeChat all count as BigData. To analyse BigData, new analytical methods have to be developed to feed business, government and organization needs.

Distributed computing and parallel processing techniques are widely used in industry for BigData applications. Hadoop (High-availability distributed object oriented platform), the most popular open-source platform for reliable, scalable, distributed computing, is often

978-0-7695-5096-1/13 $31.00 © 2013 IEEE
DOI 10.1109/CSE.2013.149

referred to by BigData researchers. Two main core frameworks in Hadoop: Hadoop Distributed File System (HDFS) and MapReduce, have being deployed in industries for the management of cluster distributed data centers such as Facebook, Google, Yahoo, Amazon.com, eBay and Twitter (hadoop.appache.org).

A. *Motivation*

The six data dimensions, or 5Ws data dimensions, (Why the data occurred, where the data came from, what the data is, how the data was transferred, who received the data and when the data occurred) for BigData analysis and visualization are not addressed by previous works, to the best of our knowledge.

Current BigData visualization approaches often reduce high dimension data to low dimension, and omit some data trends or relationships. The visual graph displays raw data through multiple clusters or multiple views using different shapes and colours to help identified data patterns. But too many lines or nodes are used to display these massive linking or nodes, which are hard for businesses, governments and organizations to understand [4][5][7].

Our approach uses the 5Ws data dimensions to establish BigData patterns to uncover the correlations which traditional database management tools could not reveal. The clustered visualization methods display the 5Ws data dimensions without any linking between data sources and data destinations.

B. *Our contributions*

In this paper, we have further developed our visual analytics model [8] and [13] for BigData analysis. The BigData pattern established can handle multiple datasets. First, we analyzed the attributes of datasets and then introduced the 5Ws data dimensions for classification. Second, we established the 5Ws model that measures data sending and receiving patterns. Third, we introduced a visualization method to illustrate BigData patterns and its evaluation for different datasets and resources.

The 5Ws model with clustered visualization provides a clear outline of data patterns that significantly change the measurement for BigData analysis and visualization. Our contributions are:

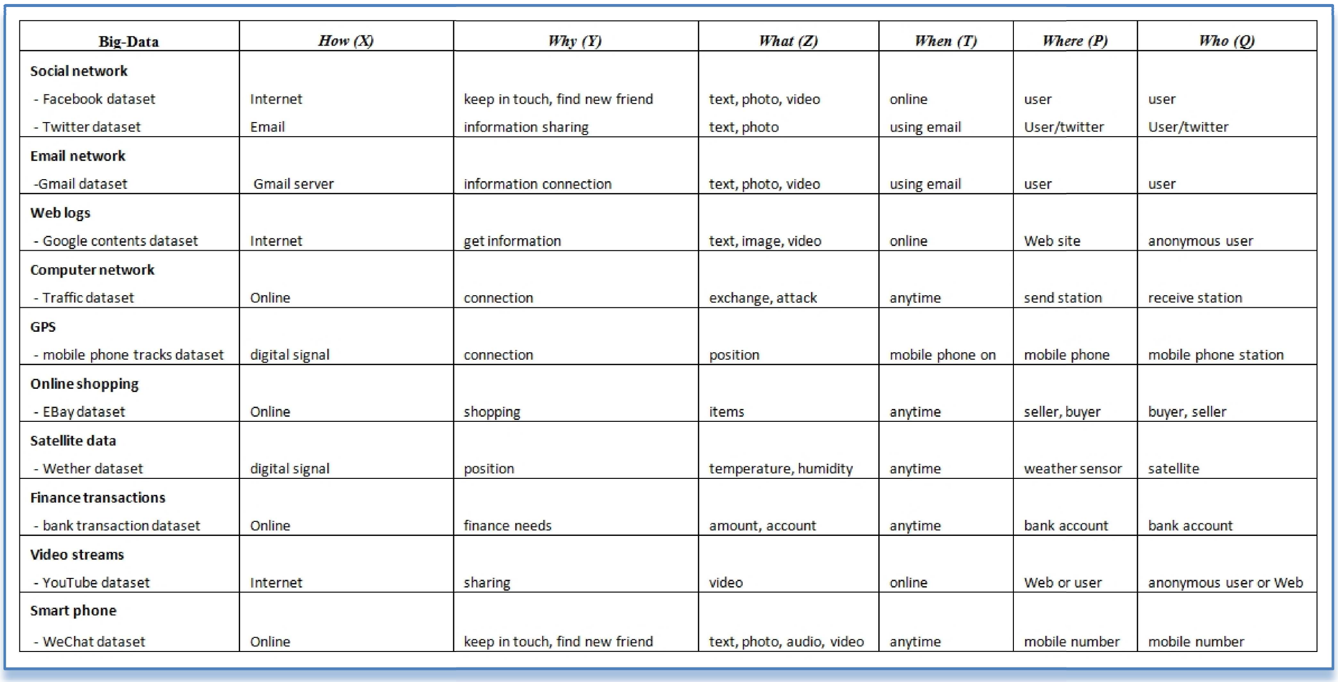
- Introduced 5Ws data dimensions to illustrate BigData attributes across different datasets
- Established density patterns based on data subset to measure BigData behaviours
- Introduced visual clustering method to display data patterns without any linking between data sources and data destinations

The paper is organized as follows; Section 2 illustrates our 5Ws data analytics model, Section 3 demonstrates the implementation and visualization, Section 4 describes related works, and Section 5 summarises our approach and future works.

II. 5Ws MODEL

A. *5Ws data dimensions*

Each data item can be classified into 5Ws data dimensions, which stands for ***what*** the data is, ***where*** the data came from, ***when*** the data occurred, ***who*** received the data, ***why*** the data occurred and ***how*** the data was transferred. Fig. 2 shows the example of BigData in the 5Ws data dimensions crossing multiple datasets and resources.



Big-Data	*How (X)*	*Why (Y)*	*What (Z)*	*When (T)*	*Where (P)*	*Who (Q)*
Social network						
- Facebook dataset	Internet	keep in touch, find new friend	text, photo, video	online	user	user
- Twitter dataset	Email	information sharing	text, photo	using email	User/twitter	User/twitter
Email network						
-Gmail dataset	Gmail server	information connection	text, photo, video	using email	user	user
Web logs						
- Google contents dataset	Internet	get information	text, image, video	online	Web site	anonymous user
Computer network						
- Traffic dataset	Online	connection	exchange, attack	anytime	send station	receive station
GPS						
- mobile phone tracks dataset	digital signal	connection	position	mobile phone on	mobile phone	mobile phone station
Online shopping						
- EBay dataset	Online	shopping	items	anytime	seller, buyer	buyer, seller
Satellite data						
- Wether dataset	digital signal	position	temperature, humidity	anytime	weather sensor	satellite
Finance transactions						
- bank transaction dataset	Online	finance needs	amount, account	anytime	bank account	bank account
Video streams						
- YouTube dataset	Internet	sharing	video	online	Web or user	anonymous user or Web
Smart phone						
- WeChat dataset	Online	keep in touch, find new friend	text, photo, audio, video	anytime	mobile number	mobile number

Figure 2. Example of BigData in 5Ws data dimensions crossing multiple datasets

We define six sets to represent the 5Ws data dimensions.

- set $T=\{t_1, t_2, t_j, \ldots, t_m\}$ representing when the data occurred
- set $X=\{x_1, x_2, x_j, \ldots, x_m\}$ representing how the data was transferred
- set $Y=\{y_1, y_2, y_j, \ldots, y_m\}$ representing why the data occurred
- set $Z=\{z_1, z_2, z_j, \ldots, z_m\}$ representing what the data is
- set $P=\{p_1, p_2, p_j, \ldots, p_m\}$ representing where the data came from
- set $Q=\{q_1, q_2, q_j, \ldots, q_m\}$ representing who received the data

Therefore, each data can be defined as a node as

$$f\,(t, x, y, z, p, q)$$

$t \mid T\{\ \}$ is the time stamp for each data incidence.

$x \mid X\{\ \}$ represents how the data was transferred, such as “by Internet”, “by email” or “online transferred”.

$y \mid Y\{\ \}$ represents why the data occurred, such as “sharing photos”, “finding new friends” or “spreading a virus”.

$z \mid Z\{\ \}$ represents what the data is, such as “video”, “image”, “text” or “number”.

$p \mid P\{\ \}$ represent where the data came from, such as “from twitter”, “smart phone” or “hacker”.

$q \mid Q\{\ \}$ represent who received the data, such as “friend”, “bank account” or “victim”.

All data in the T time slot, represent as a set F with a number n incidences, is defined as

$$F = \{f_1, f_2, f_3, \ldots, f_n\} \tag{1}$$

F contains all incident nodes within a certain time period. For example, there were 9.66 million tweets during the Opening Ceremony of the London 2012 Olympic Games [1]. The twitter dataset for the Opening Ceremony is therefore $|F| = 9.66$ million.

For a particular attribute node where $x=\alpha$, $y=\beta$, $z=\gamma$, $p=\delta$ and $q=\varepsilon$, the node can then be defined as

$$f_{(\alpha, \beta, \gamma, \delta, \varepsilon)} = f\,(t, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p_{(\delta)}, p_{(\varepsilon)}) \tag{2}$$

A subset $F_{(\alpha, \beta, \gamma, \delta, \varepsilon)}$ that contains all the particular attributed nodes $f_{(\alpha, \beta, \gamma, \delta, \varepsilon)}$ in the T time slot is therefore defined as

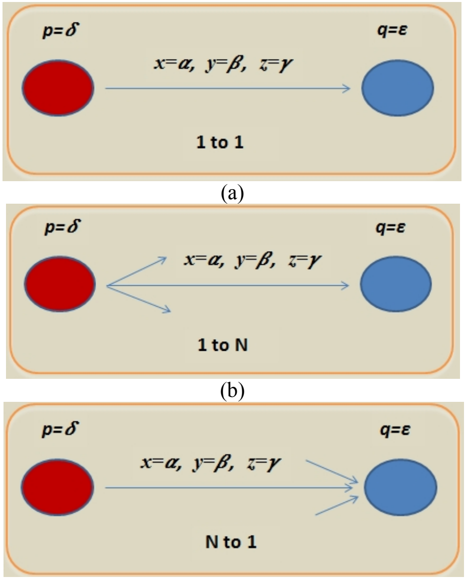
$$F_{(\alpha, \beta, \gamma, \delta, \varepsilon)} = \{\, f \in F \mid f\,(t, x, y, z, p, q),\ x=\alpha,\ y=\beta, z=\gamma, p=\delta,\ q=\varepsilon \,\} \tag{3}$$

The subset $F_{(\alpha, \beta, \gamma, \delta, \varepsilon)}$ represents the particular incident nodes by the 5Ws data dimensions. For example, during the Opening Ceremony of the London 2012 Olympic Games, 9.66 million tweets contain multiple patterns such as α = “sent or received”, β =“sharing opening ceremony” or “enjoying ceremony”, γ = “London” + “Olympics” + “Opening” + “Ceremony” plus more, δ = twitter, ε = users and t = 27-Jul-2012, 21:00 – 00:45.

The datasets $|F|$ illustrates the statistical results in volume and velocity. The subset $F_{(\alpha, \beta, \gamma, \delta, \varepsilon)}$ demonstrates the variety for the particular incident pattern, which provides more analytical features for business, government and organizations.

B. Data transfer patterns

The subset $F_{(\alpha, \beta, \gamma, \delta, \varepsilon)}$ contains information about where the data came from (sender δ), who received the data (receiver ε) and how the data was transferred (α). The sender (δ) and receiver (ε) can be a person, location, system, or any attributes that sent to received data. Three basic data transfer patterns are shown in Fig. 3.



(a)

(b)

(c)

Figure 3. Three basic data transfer patterns

(a) represents a data transfer pattern of 1:1, which means that the data occurred only between the sender (δ) and the receiver (ε). If the pattern is 1: N, as shown in (b), it indicates that the sender (δ) sent multiple data and (ε) is one of the receivers. A pattern of N: 1 is shown in (c), which indicates that multiple data is sent to the receiver (ε) and (δ) is one of the senders. The pattern N:N can be described as a combination of these three basic patterns, N:N = (1:1) + (1:N) + (N:1).

C. Data sending density

We introduced sending density (SD) to measure the sender’s pattern during data transferal. Based on (3), the sending density for particular attributes $x=\alpha$, $y=\beta$, $z=\gamma$, and $p=\delta$, in time slot T, is defined as $SD_{(\alpha, \beta, \gamma, \delta)}$

$$SD_{(\alpha, \beta, \gamma, \delta)} = \frac{|\,F(\alpha,\beta,\gamma,\delta)\,|}{|\,F\,|} \tag{4}$$

$$= \frac{1}{n} \sum_{i=1}^{n} f_{(i)}\big(t, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p_{(\delta)}, q\big)$$

where $0 \leq SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} \leq 1$

$SD_{(\alpha,\, \beta,\, \gamma,\, \delta)}$ represents the 5Ws dimensions for the sender's pattern; the content was γ, transferred by α, in $t \subset T$ time, for reason β and sent by δ. A high value of $SD_{(\,)}$ indicates where the most data came from. Fig 3 (b) illustrates the example of a high value of $SD_{(\,)}$.

D. Data receiving density

The receiving density for $x = \alpha$, $y = \beta$, $z = \gamma$ and $q = \varepsilon$, is defined as $RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$ as

$$RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)} = \frac{|\, F(\alpha,\beta,\gamma,\varepsilon) |}{|\, F\, |}$$

$$= \frac{1}{n} \sum_{i=1}^{n} f_{(i)}\big(t, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p, q_{(\varepsilon)}\big) \qquad (5)$$

where $0 \leq RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)} \leq 1$

$RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$ represents the 5Ws dimensions for the receiver's pattern; the contents was γ, transferred by α, in $t \subset T$ time, for reason β and received by ε. A high value of $RD_{(\,)}$ indicates who received the most data. Fig 3 (c) displays the example of a high value of $RD_{(\,)}$.

E. Density cross datasets

$SD_{(\,)}$ and $RD_{(\,)}$ not only measure the patterns for one dataset, but can also be used to compare multiple datasets. For example, a Facebook dataset and a bank transaction dataset are two different datasets. But similar attributes exist on both datasets, such as δ = "users", α = "mobile connection", the comparison between those two datasets for δ = "users". α = "mobile connection" will export the ratio of internet banking mobile users via Facebook mobile users. Those two densities provide the more measurement features for BigData analysis.

F. Density classification

When $SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} = RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$, it indicates that the data transferred pattern is 1:1, shown as Fig 3 (a). If $SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} > RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$, it represents that data transferred pattern is 1:N, shown as Fig 3 (b), and for N:1, $SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} < RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$, as shown as Fig 3 (c).

Here, we introduce a coefficient $\theta_{(\,)}$ to classify the density patterns for the attributes α, β, γ.

$$\theta_{(\alpha,\, \beta,\, \gamma)} = \frac{RD()}{SD()}$$

$$= \frac{\sum_{i=1}^{n} f_{(i)}\big(t, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p, q_{(\varepsilon)}\big)}{\sum_{i=1}^{n} f_{(i)}\big(t, x_{(\alpha)}, y_{(\beta)}, z_{(\gamma)}, p_{(\delta)}, q\big)} \qquad (6)$$

When $SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} = RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$, $\theta_{(\,)} = 1$. If $SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} < RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$, then $\theta_{(\,)} > 1$, and $\theta_{(\,)} \rightarrow \infty$ if $SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} \rightarrow 0$. When $SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} > RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)}$, then $\theta_{(\,)} < 1$, and $\theta_{(\,)} \rightarrow 0$ if $RD_{(\alpha,\, \beta,\, \gamma,\, \varepsilon)} \rightarrow 0$. Fig. 4 shows an example of density patterns for $SD_{(\,)}$ and $RD_{(\,)}$.

Three curves (green, red and orange) are examples of the different patterns of $SD_{(\,)}$ and $RD_{(\,)}$. The yellow area represents $SD_{(\,)} < RD_{(\,)}$ and $\theta_{(\,)} > 1$. The grey area represents $SD_{(\,)} > RD_{(\,)}$ and $\theta_{(\,)} < 1$.

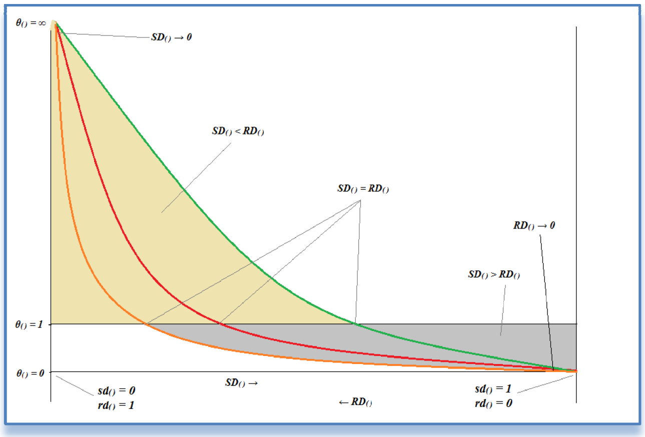


Figure 4. Example of density patterns

When $\theta_{(\,)} \rightarrow \infty$ and $RD_{(\,)} \rightarrow 1$, it indicates the value of $SD_{(\,)}$ is very low. This shows that huge amounts of data with the same attributes (α, β, γ) were sent to a receiver by multiple senders. When $\theta_{(\,)} \rightarrow 0$ and $SD_{(\,)} = 1$, it means $RD_{(\,)}$ is very low, which demonstrates that huge amounts of data with the same attributes (α, β, γ) were sent from a sender to multiple receivers.

The density pattern $\theta_{(\,)}$ can be used for many purposes. Back to the example of having 9.66 million tweets during the Opening Ceremony of the London 2012 Olympic Games. Suppose α = "iPhone apps", β ="sharing photo", γ = "James Bond" + "Queen" + "parachute", δ = "users" and ε = "Twitter receiving server in London". The value of $\theta_{(\,)} >> 1$ because huge amount of text messages and images were sent to one receiver from multiple iPhones. $\theta_{(\,)}$ can also be used for comparing different attributes such as α_1 = "iPhone apps" and α_2 = "Android apps", or γ_1 = "USA team" and γ_2 = "UK team".

G. Noise data calculation

The noise data is defined as the unknown nodes in the density algorithm methods, such as $x = unknown\_x$, $y = unknown\_y$, $z = unknown\_z$, $p = unknown\_p$ and $q = unknown\_q$. A subset for unknown nodes can be defined as

$$\begin{aligned} F_{(unknown\,)} = \{ f \in F \mid f(t, x, y, z, p, q),\ & x = unknown\_x \\ \square\ y = unknown\_y\ \square\ & z = unknown\_z \\ \square\ p = unknown\_p\ \square\ & q = unknown\_q \,\} \end{aligned} \qquad (7)$$

(4) would be amended as

$$SD_{(\alpha,\, \beta,\, \gamma,\, \delta)} = \frac{|\, F(\alpha,\beta,\gamma,\delta) |}{|\, F\, | - |F(unknown\,)|} \qquad (8)$$

(5) would be amended as

$$RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)} = \frac{|\ F(\alpha,\beta,\gamma,\varepsilon)|}{|\ F\ |-|F(unknown\ )|} \qquad (9)$$

$SD_{(\ )}$ and $RD_{(\ )}$ represents the sender's and receiver's pattern, which significantly improves the accuracy of BigData analysis, because both densities avoid noise data.

H. 5Ws clustered visualization

Fig. 5 shows an example of 5Ws tree map. The Tree map is widely used to illustrate data structures and its contents. Hundreds, even thousands, of different attributes can be classified by our 5Ws data dimensions tree.

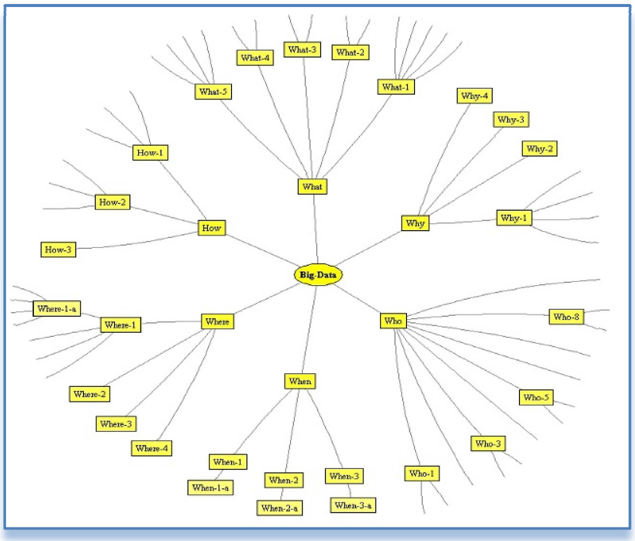


Figure 5. 5Ws data dimension tree

Clustering is the process of organizing data into groups with similar elements in some attributes [6]. Each data represents a node that contained the 5Ws data dimensions for each classification. After gained values of $RD_{(\ )}$ and $SD_{(\ )}$, we use the five clustering levels to illustrate both densities. The first clustered level is x=α, the second is z=γ, the third is y=β, and fourth and fifth are p=δ and q=ε.

We use a visual circle to represent the attributes and its patterns, with points in the graph. We do not show any linking because the data patterns have been calculated before the clustering visualization. Accordingly, the radius of the circle for a particular attribute is calculated by its density, which is defined as

$$R_{(\alpha)} = \sqrt{\frac{|\ F\ (\alpha)|}{|\ F\ |}}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_{(i)}\big(t, x_{(\alpha)}, y, z, p, q\big)} \qquad (10)$$

where $R_{(\alpha)}$ is the radius for x=α.

$$R_{(\beta)} = \sqrt{\frac{|\ F\ (\beta)|}{|\ F\ |}}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_{(i)}\big(t, x, y_{(\beta)}, z, p, q\big)} \qquad (11)$$

where $R_{(\beta)}$ is the radius for y= β.

$$R_{(\gamma)} = \sqrt{\frac{|\ F\ (\gamma)|}{|\ F\ |}}$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n} f_{(i)}\big(t, x, y, z_{(\gamma)}, p, q\big)} \qquad (12)$$

where $R_{(\gamma)}$ is the radius for z= γ.

p=δ and q=ε are not calculated for the radius because it takes too much space to display it in the graph. But, p and q will be displayed separately after the clustering visualization is finalized, which is shown in the next sector.

III. IMPLEMENTATION AND VISUALIZATION

We have tested the 5Ws model by using the network security dataset, ISCX2012 dataset [9], which contains 20 attributes shown in Appendix. Table II displays the summary of one ISCX2012 dataset.

TABLE II. ISCX2012 dataset - TestbedTueJun15c

Name	Amount
Network traffic nodes	130288
Source Ips	36
Destination IPs	1656
ICMP traffics	31
TCP traffics	119242
Unknown TCP traffics	3
UDP traffics	11015
Unknown UDP traffics	36
Connecting methods	19
Source ports	23653
Destination ports	222
Attacks	37375

Assume that α = "TCP traffic", β = "HTTP connection", γ = "Attack", δ = "Source IP and port" and ε = "Destination IP and port" for this simulation. The result has illustrated that the destination IP address ε = 192.168.5.122:80 is targeted as the victim, and faced γ = *"attack"* by α = *"TCP"* traffic with β = *"HTTP"* connection. There are six different sources IPs (δ) that sent the attacks to the victim. The value of six different $SD_{(\alpha,\ \beta,\ \gamma,\ \delta)}$ and one $RD_{(\alpha,\ \beta,\ \gamma,\ \varepsilon)}$ are shown as bubbles in Fig. 6. The six different density patterns, represented as the values of $\theta_{(\alpha,\ \beta,\ \gamma)}$, are shown in Fig. 7.

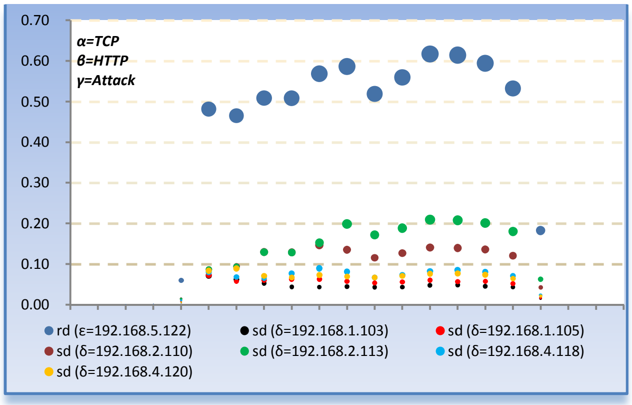


Figure 6. Values of $sd_{()}$ and $rd_{()}$

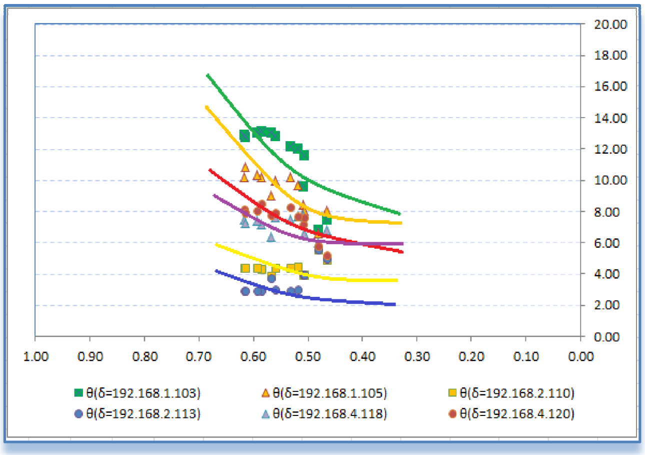


Figure 7. values of $\theta_{(\alpha,\ \beta,\ \gamma)}$

In simulation, the values of $SD_{(\ )}$ and $RD_{(\ )}$ start increasing from 16:05, demonstrated in Fig 6. During the threat, both $SD_{(\ )}$ and $RD_{(\ )}$ are kept at high levels until 17:20. It gives the alert for the network intrusion detection at an early stage. The density patterns are illustrated in Fig. 7, which show that the hacker sent different attack patterns from six sources IPs.

Fig. 8 shows that the visual nodes in random spots before clustering visualization. The clustered structure is dependent on the values of $SD_{(\ )}$ and $RD_{(\ )}$, which are represented as yellow nodes. The different attributes of $SD_{()}$ and $RD_{(\ )}$ use different scales to save space. For example, $SD_{(3)}$ may represent the value of FTP connection, and $RD_{(6)}$ may indicate the value of ICMP traffic. Before clustering progress, there is no linking edge connected to any node.

We preset $SD_{(\ )}$=0.80 and $RD_{(\ )}$=0.80 to start the visualization process, but unfortunately, no clustered structure have appeared. This is because the value of $SD_{()}$ and $RD_{()}$ are not reached at these points. The values of $SD_{()}$ and $RD_{(\ )}$ were decreased until $RD_{(\ )}$ = 0.50 and correspondingly $SD_{(\ )}$ = 0.10. The final clustered structure generated is shown in Fig. 9.

The two top values of $SD_{()}$ nodes and one top value $RD_{(\ )}$ node appeared in the final clustered visualization. The attribute x=α is the first clustered level for visualization. z=γ is the second and y=β is the third. All nodes that have the same classifications for $RD_{(k)}$ and $SD_{(j)}$ have been gathered into groups, linked by the different clustered level (k and j is used for separating them from other nodes). p and q has been displayed at the bottom field which illustrates where the data came from and who received the data.

The nodes that do not belong to $RD_{(k)}$ and $SD_{(j)}$ are still shown in the graph as random points. The clustering visualization scales down the density patterns into the particular attributes. This enables the network intrusion detection to be seen easily, efficiently and clearly.

Our clustered visualization has avoided the mass linking crosses that the nodes provided, therefore showing a clear structure for the pattern detection. The higher value of $SD_{()}$ indicates that the hacker sent widespread attack traffic to the victim's systems across the network. The high value of $RD_{()}$ means that the victim suffered the flood attacks from multiple sources.

Therefore, the 5Ws model with clustered visualization has provided clear outlines of data patterns, which significantly change the measurement of BigData analysis and visualization.

IV. RELATED WORKS

The researchers have practiced HDFS and MapReduce for data processing, data sharing or data clustering because it has the ability to take a query over a dataset, divide it into many small fragments, and run it in parallel to the cluster. In addition, it provides a distributed system that stores data on the computer nodes, with high aggregate bandwidth across the cluster [6][10][11][12].

SAS Visual Analytics Explorer [4] scales, visualizes and analyse the massive dataset to find visual patterns. Heat maps and tree maps were used in the visualization. An example of a head map for BankWorld's activity is shown in Fig. 10. They created the head map based on the latitude and longitude of each log entry. The raw data is displayed in the graph.

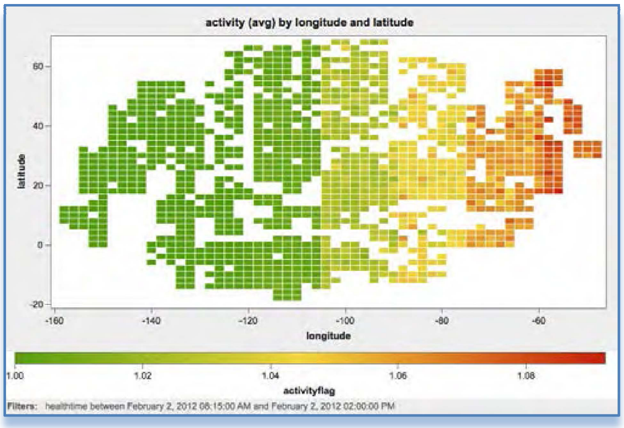


Figure 8. Heat map of BankWorld's activity [4]

Cheng-Long Ma et al [5] used the K-means clustering method to find out the clustering centers for 3-D visualization. The distances of the three coordinate axes are corresponded by data in the original space. They tested their model using the Iris database, as shown in Fig. 11.

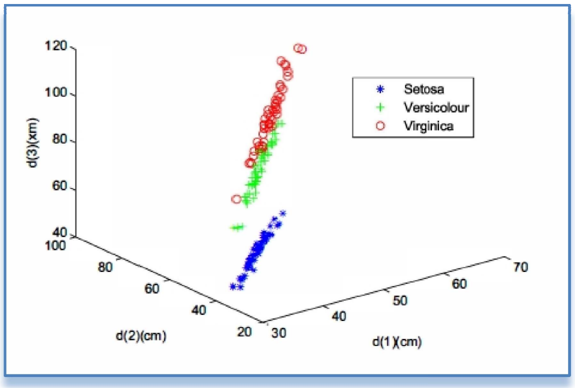


Figure 9. Iris 3-D visualization [5]

Unfortunately, current visualization approaches cannot display BigData patterns crossing multiple datasets, as too many lines or nodes are used in the graph. We couldn't make many comparisons because there is, to the best our knowledge, no previous work which addresses the 5Ws data dimensions for BigData analysis and visualization.

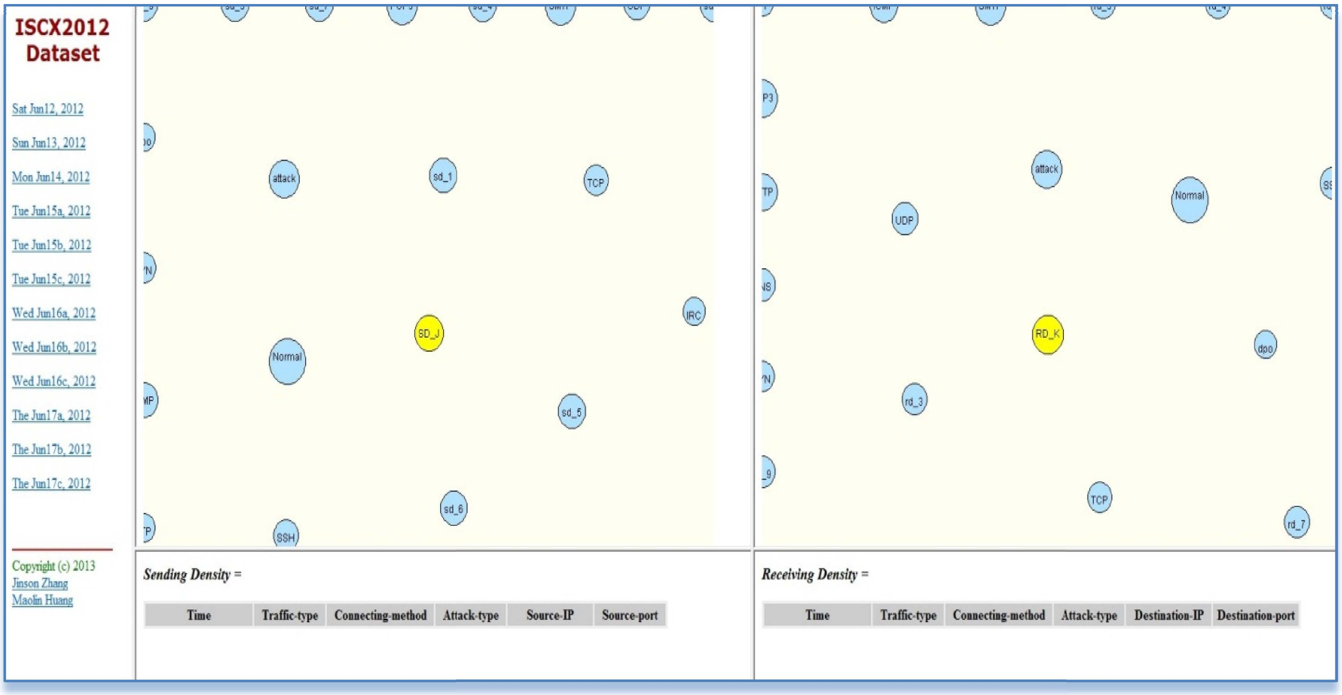
V. CONCLUSIONS & FUTURE WORK

In this paper, we have established the 5Ws model, the framework for BigData analysis and visualization. The model analyses the 5Ws data dimensions crossing multiple datasets and builds sending density and receiving density to measure BigData patterns. This is done to uncover any correlations which traditional database management tools could not revealed.

5Ws model not only measures BigData patterns, but also enables density comparisons between multiple datasets. This provides more analytical features of BigData analysis for business, government and organizational needs.

Our clustered visualization method displays the 5Ws data dimensions without any linking between data sources and destinations. This allows users to interactively select and scale down views of BigData patterns. The experiment shows that this model, with the clustered visualization, can be used effectively for BigData analysis and visualization.

For the future work, we plan to develop our 5Ws model in three directions. Firstly, we plan to deploy the densities classification in more areas. Secondly, we plan to apply our 5Ws model for more datasets. Thirdly, we plan to use the Gapminder's visualization technique developed by Dr. Hans Rosling (www.gapminder.org), the world famous visualization presenter, to display 5D data dimensions through 2D graphs.



ISCX2012 Dataset

Sat Jun12, 2012

Sun Jun13, 2012

Mon Jun14, 2012

Tue Jun15a, 2012

Tue Jun15b, 2012

Tue Jun15c, 2012

Wed Jun16a, 2012

Wed Jun16b, 2012

Wed Jun16c, 2012

The Jun17a, 2012

The Jun17b, 2012

The Jun17c, 2012

Copyright (c) 2013
Jinson Zhang
Maolin Huang

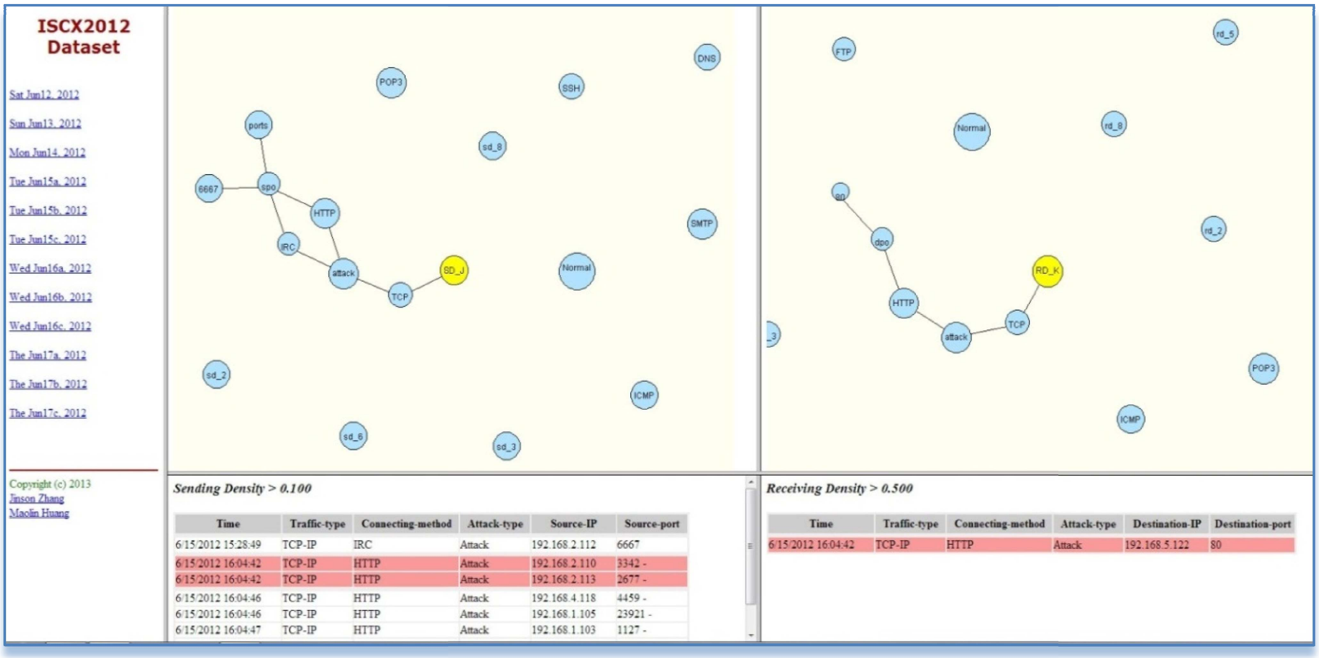
Sending Density =

Time	Traffic-type	Connecting-method	Attack-type	Source-IP	Source-port

Receiving Density =

Time	Traffic-type	Connecting-method	Attack-type	Destination-IP	Destination-port

Figure 10. The nodes before clustering process



Sending Density > 0.100

Time	Traffic-type	Connecting-method	Attack-type	Source-IP	Source-port
6/15/2012 15:28:49	TCP-IP	IRC	Attack	192.168.2.112	6667
6/15/2012 16:04:42	TCP-IP	HTTP	Attack	192.168.2.110	3342 -
6/15/2012 16:04:42	TCP-IP	HTTP	Attack	192.168.2.113	2677 -
6/15/2012 16:04:46	TCP-IP	HTTP	Attack	192.168.4.118	4459 -
6/15/2012 16:04:46	TCP-IP	HTTP	Attack	192.168.1.105	23921 -
6/15/2012 16:04:47	TCP-IP	HTTP	Attack	192.168.1.103	1127 -

Receiving Density > 0.500

Time	Traffic-type	Connecting-method	Attack-type	Destination-IP	Destination-port
6/15/2012 16:04:42	TCP-IP	HTTP	Attack	192.168.5.122	80

Figure 11. The higher value of $SD_{()}$ and $RD_{()}$ after clustering visualization

REFERENCES

[1] Pingdom, “Internet 2012 in numbers”, posted on Jan 16, 2013, http://royal.pingdom.com/2013/01/16/internet-2012-in-numbers/

[2] Stamford, “Gartner Says Solving ‘Big Data’ Challenge Involcves More Than Just Managing Volumes of Data”, posted on June 27, 2011, http://www.gartner.com/newsroom/id/1731916

[3] D. Klein, P. Tran-Gia, M. Hartmann “Big Data”, Informatik-Spektrum, vol 36, issue 3, pp319-323, June 2013

[4] N.A. Abousalh-Neto, and S. Kazgan, “Big Data Exploration through Visual Analytics”, In Proc. IEEE Symposium on Visual Analytics Science and Technology 2012, pp 285-286, Oct 2012

[5] C.L. Ma, X.F Shang, and Y.B Yuan “A Three-Dimensional Display for Big Data Sets”, In Proc. 2012 International Conference on Machine Learning and Cybernetics, pp 1541-1545, July 2012

[6] J. Hurwitz, A. Nugent, F. Halper, M. Kaufman, Big Data for Dummies, Published by John Wiley & Sons. Inc, 2013

[7] J. Choo, H. Park, “Customizing Computational Methods for Visual Analytics with Big Data”, IEEE Computer Graphics and Applications, vol 33, No 4, pp 22-28, July 2013

[8] J. Zhang, and M.L. Huang, “Visual Analytics Model for Intrusion Detection in Flood Attack”. In Proc. TrustCom 2013, 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp 277-284, July 2013

[9] A. Shiravi, H. Shiravi, M. Tavallaee, and A.A. Ghorbani, “Toward developing a systematic approach to generate benchmark datasets for intrusion detection,” Computers & Security, vol 31, issue 3, May 2012, pp 357-374, ISSN 0167-4048

[10] T. Kraska, "Finding the Needle in the Big Data Systems Haystack," Internet Computing, IEEE, vol.17, no.1, pp 84,86, Jan-Feb. 2013

[11] S. Narayan, S. Bailey, A. Daga,"Hadoop Acceleration in an OpenFlow-Based Cluster," High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:, pp 535,538, 10-16 Nov. 2012

[12] A. Menon, “Big Data @ Facebook”, In Proc. 2012 workshop on management of big data system, pp 31-32, 2012

[13] J. Zhang, M.L. Huang and D. Hoang, “Visual analytics for intrusion detection in spam emails”, International Journal of Grid and Utility Computing, vol 4, no 2/3, pp 178-186, 2013

APPENDIX

ISCX2012 dataset contains 20 attributes. The example of attributes in 5Ws dimensions is shown as below.

When (T)	StartDateTime, StopDateTime
How (X)	ProtocolName, Direction
Why (Y)	AppName, SourceTCPFlagsDescription, DestinationTCPFlagsDescription
What (Z)	Tag, SourcePayloadAsBase64, SourcePayloadAsUTF, DestinationPayloadAsBase64, DestinationPayloadAsUTF
Where (P)	Source, SourcePort, TotalSourceBytes, TotalSourcePackets,
Who (Q)	Destination, DestinationPort, TotalDestinationBytes, TotalDestinationPackets,