Big Data Real-Time Storytelling with Self-service Visualization

Rajat Kumar Behera and Anil Kumar Swain

Abstract Stories help to communicate information and interpret knowledge. Once the data is collected, analyzed, cleansed, and transformed, the subsequent step is to extract potential value from it. Realization of value will happen, only when business-centric insights are discovered and translated to time-bound actionable outcome. To maximize the potential value, data should be decoded into a storytelling medium via visualization, which can be either static or dynamic. Big data visualization is to reveal stories from data tsunami, generated at an alarming speed with diversified formats. The stories tend to represent vital characteristics to enlarge users. Self-service visualization empowers users to uncover unique patterns, interesting facts, and relationships from the underlying data by building their own stories without the in-depth technical knowledge, possibly little handhold by IT department. In this survey paper, we first get familiar with big data storytelling with visualization and its related concepts, and then will look through general approaches to do the visualization. To get deeper about it, we will have discussion about truthful data visualization in self-service mode representing real view of the business. This paper also presents the challenges and available technological solution, covering open source for representing real-time view of the story.

1 Introduction

Big data is exemplified by sheer volume, soaring velocity, diversified variety, and/or inherent discrepancies, i.e., veracity (4 Vs) datasets that use a new paradigm of processing for insight discovery and time-bound decision-making [1]. Such data is generating unparalleled options for the businesses to culminate deeper insights to

R. K. Behera (✉) · A. K. Swain
KIIT Deemed to be University, Bhubaneswar, India
e-mail: rajat_behera@yahoo.com

A. K. Swain
e-mail: anilkumarswain@gmail.com

A. Abraham et al. (eds.), *Emerging Technologies in Data Mining and Information Security*, Advances in Intelligent Systems and Computing 813,
https://doi.org/10.1007/978-981-13-1498-8_36

Table 1 Benefits and time requirement by industry

Industry	Time Req	Expected benefits
Clinical care	Seconds	Reduce life risks and saves lives
Financial and stock market	Milliseconds	Enhance performance and business profit
Military decision-making	Seconds	Saves lives and enhance better performance
Intelligent transportation	Seconds	Save times and enhance living quality
Natural disasters	Minutes	Reduces life risks
Festivals/Crowd control	Seconds	Efficient crowd handling
Daily resources	Minutes	Efficient resource administration

Table 2 Benefits of data visualization technique

Sr#	Benefits	Percentages
1	Enhanced decision-making	77
2	Enhanced ad hoc data analysis	43
3	Enhanced collaboration/information sharing	41
4	Afford self-service capabilities to end users	36
5	Better return on investment	34
6	Time savings	20

reinforce the decision-making process and leads the struggle for not only to store and process the data in a significant way, but to present it in meaningful ways.

Different applications in real-time big data analytic can be used in many aspects of human life like quality of life, minimization of risks of lives, resource management efficiency and profitability in business, etc. So, we need to analyze and execute big data in real-time analytic field as soon as possible to get a fast response [2]. Real-time big data application related to industry like financial and stock market, intelligent transportations, early warning of natural disasters, etc., are important applications whose operations help in enhancing quality of life, reducing human risks and saving the lives of people. Due to real-time requirements, many challenges give more attention to collecting, transferring, processing and visualizing big data [3]. Table 1 is depicting the benefits and time requirement by respective industry [2].

To compete more efficiently and effectively, businesses are increasingly turning to data visualization technique that allows the decision maker to visualize a custom analytic view. Data visualization is the technique to represent the data, including variables and attributes in an orderly form for the unit of information visualization [4]. Visualization can be thought of as the "face" of big data. According to the respondent percentage of survey [5], Table 2 represents the benefits of data visualization.

The self-service aspect of the visualization allows analyst to create visualizations at their own, in their own time, while still matching the functionality and capabilities of non-self-service aspect. It is not merely the cost and efficiency savings that come as a by-service of self-service methods. With the right self-service data visualization tools and software, analysts are able to uncover interesting and unique patterns and relationships from the underlying data, and create striking visualizations to express those patterns through powerful, memorable visuals.

In an age where big data along with social, mobile and the cloud are all converging in new and exciting ways, data storytelling has become more essential than ever. Big data storytelling is a technique of delivering information resulting from multifaceted data analysis process in a way that allows the decision-makers to easily and quickly understand the context, understand its meaning and draw conclusions from it [6]. The vital step in constructing any storytelling is structuring easy-to-follow data stories which are the sequences of causally related events. Foremost, storytelling takes time to unfurl and the tempo matches the audience's aptitude to follow. Next, it holds the audience's attention by encompassing attention-grabbing background, typeset, and intrigue. Finally, it leaves a eternal perception, either by stimulating the audience's curiosity and making them want to study or observe more.

2 Preliminaries and Basic Concepts

In this section, we present the definitions and basic concepts of big data real-time analysis, real-time self-service visualization, and real-time data storytelling.

2.1 Big Data Analysis in Real Time

To perform real-time big data analysis, three major steps are required and is presented in Fig. 1.

Each of the steps is described below covering technical challenges:

1. Real-Time Data Collection—It is the method of collecting data from diversified sources and then integrates and stores in as-is form. The challenge involved in establishing trusted connection to various data sources and extracting and storing in real-time.
2. Real-Time Data Processing—The process essentially gives shape to stored raw data. Precisely, it cleanses and transforms for visualization or analysis consumption. Additionally, it produces smaller datasets (with aggregation/summarization

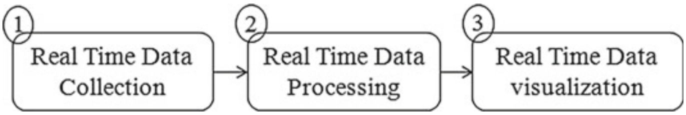

1
Real Time Data Collection
2
Real Time Data Processing
3
Real Time Data visualization

Fig. 1 Real-time data big data analysis steps

operation) which are generally consumed for visualization. It is also designed to detect real-time business in the processed data and to raise early warning for immediate action. Detection of exception is generally predefined business rules. The challenges surfaced while querying the large amount of data for abnormality detection.

3. Real-Time Data Visualization—This process involves monitoring the actions defined by the real-time decision-making process. The challenges surfaced when the visualization are explicitly performed by human in manually rather relying on available visualization tools.

2.2 Big Data Self-Service Visualization

Data visualizations are the method for expressing the insights found within big data. Self-service tools let the analyst without technical experience or know-how can make use of the tool without the need for external assistance to represent the insights from the underlying massive volume of data. It is added level of control, flexibility and customization over the work. Real-time self-service visualization allows making data-driven decisions with no gaps in information.

Traditional big data visualization development was confined to few technical professionals and few decision-makers were able to access the data. This approach was analogous to the conventional top-down business ladder where only a few decision-takers were making the decisions. This classical decision-making process does not fit modern era. A culture of bottom-up is getting traction for decision-making and enables everyone to visualize the data and analysis and has the potential to make even better data-driven decisions. In traditional visualization, the analyst and end user typically have no ability to create their own data visualizations or established data sources without the assistance of technical professionals. Self-service visualization empowers the analyst to explore any reachable dataset and to easily create personalized data visualization. Typically, it does not require a traditional deployment methodology such as build, test, and publish, since every analyst can simply build or extend the visualization with graceful methods such as drag-n-drop.

The difference between traditional visualization and real-time self-service visualization is presented in Table 3.

2.3 Real-Time Data Storytelling

Traditional visualization results in protracted visualization with over-eventful text, tables, graphs, and other visualization components. Decision-takers then regularly kick off exercise to reduce the mountain of visualization, demanding for summary visualization or exception visualization. The essence is to ensure that all the pertinent

Table 3 Difference between visualization and self-service visualization

Traditional visualization	Self-service visualization
Concept is old and exists around last 15 years	New concept, getting traction
IT department create visualization for each analyst/end user	Analyst creates visualization and answer questions for themselves, rather than having an IT department create visualization for each question/end user request
For each new question/end user ad hoc request, it places a burden on IT and goes through traditional build-n-deploy and it consumes time	For each new question/end user ad hoc request, analyst can build of its own and leads to quick turnaround time
Has a predefined view of the data	Operations like aggregation, summing are performed on the fly and hence the real-time view
Dimensions are decided at the time of designing the view by the IT department	Each end user may be asking different questions, and looking for different answers. So on-demand dimensions can be designed seamlessly
Generally not exploit In-Memory	In-memory instead of disk for any operation
Minimal support on "what-if" analysis	Enrich with real-time "what-if" analyses and leads to smarter decision

information still persists in the condensed version and the message is easily understandable. Frequently, the condensed version does not answer the vital question.

Real-time storytelling promptly provides the context, relevance, expectations and enables the decision-makers to grapple a vast quantity of facts swiftly and is combined with the power of efficient and effective real-time visualization.

3 Literature Review

In this section, we present the literature review of "big data solution stack to overcome technical challenge" and "history of storytelling in big data visualization".

3.1 Big Data Solution Stack

Big data analysis requires the use of best technologies at every step be at collating data, processing, and finally deriving the final conclusion in the form of visualization. The best technology helps in improving the processing power of the overall system for analyzing in real time. So each of the steps needs the implementation is not just only efficient but also economical. One of the ways to achieve the combination of efficiency and economy is open-stack technology with the adoption of parallel

Table 4 Big data analysis tool

Tool	Purpose
Apache Mahout	Machine learning and Data Mining
R	Predictive/Descriptive Analytics

and distributed approach. The solution stack is addressing the following technical challenges:

1. Real-Time Data Collection—Hadoop Distributed File System (HDFS) is a distributed file system and runs on commodity hardware [7] and is based on master–slave architecture. HDFS cluster comprises a single name node and more no of data nodes. Name node manages the file system namespace and regulates access to files by client requests whereas data nodes, usually one per node in the cluster, manage storage attached to the nodes that they run on. In HDFS, a large single file is split into sizable no. of blocks, which are stored in a set of data nodes. The name node executes different operations like opening, closing, and renaming files and directories related to file system and also determines the mapping of blocks to data nodes. The data nodes are used to execute when a client raises read and write requests. The data nodes perform different operations like block creation, deletion, and replication after getting instruction from name node [7]. Apache Flume and Sqoop are both eco-component of Hadoop which pull the data and load them into Hadoop cluster. Flume is responsible for collecting and aggregating large amount of log data whereas Sqoop is responsible for collecting, storing, and processing large amount of data. It scales data horizontally and multiple Flume tools cluster should put into action to collect large amount of data parallely and storing to staging area parallely as well [8].
2. Real-Time Data Processing and Data Storage—Spark is Apache Software Foundation tool for speeding up Hadoop data processing. The main feature of Spark is its in-memory computation that guarantees increase in the processing speed by caching the data and results in real-time processing. It provides fault tolerance through RDD (Resilient Distributed Dataset) which transparent data storage on memory and persists to disc only when needed. This helps to reduce most of the disc read and write cycle [9]. A collection of spark tool does parallel read and on the fly, dirty data is discarded and only the meaningful data to be transformed and loaded to Data Lake (No SQL).
3. Real-Time Data Visualization—Following tools are adequate for real-time analysis.

    3.1 Big Data Analysis Tool—The purpose of the big data analysis tool is to detect patterns, identify trends, and collect other valuable findings from the ocean of information. Table 4 depicts the tools to perform various analytics.

    3.2 BI Tool—BI (Business intelligence) tools are the technology that enables business/operational people to visualize the information to help/better the business/operation. Tableau is an in-memory BI tool for visually analyzing the data in real time. Tableau also enables users to share interactive and shareable dashboards depicting trends, variations, and density of the data in form of graphs and charts [10].

3.2 History of Storytelling in Big Data Visualization

In 1977, John Tukey wrote in Exploratory Data Analysis that Visualization is a means to extend storytelling and communicate patterns and trends in the visual realm. The role of visualization in data science and more so in big data science is inestimable [11].

In 1997, Behrens commented on Tukey's work saying "often likened EDA to detective work". The role of the data analyst is to listen to the data in as many ways as possible until a plausible story of the data is apparent [12]. Tukey's influence on the data science community remains to this day. Ultimately, the goal of visualization is to communicate answers to a question. The authors developed a generalized data science pipeline paralleling the elements of a story with a beginning, middle, and an end.

In 2014, Blue Hill Research has adapted Campbell's approach to analytics as shown in Fig. 2 [11].

In 2014, Mico Yuk and Stephanie Diamond defined an easy-to-follow storyboard. They defined storyboarding is to translate business requirements into a four segments that states the goal, measurements, and data visualization types [13]. The segments are

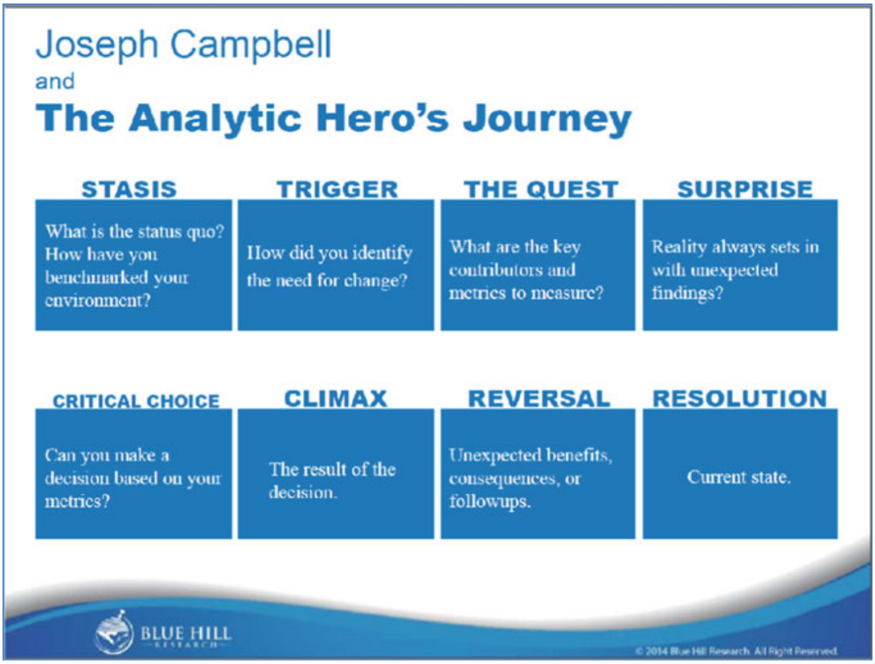
- Current State: What is happening now?
- Trends: How did it happen and what are the relationships?
- Forecast: What will happen in the future?
- What-if: How should we act in future to achieve or exceed the goals?

There are many variations to this approach. In the simplest case, a story has a beginning, middle, and an end.

In 2015, Linderman formulates five rules of storytelling [14] and is as follows.

- Make Opening Count—Explain the problem and hint the solution in first few lines.
- Be Vulnerable—Represent the challenges.
- Build Tension—Pitches are the subtle changes in the emotion. It is vital to identify the arc of the pitch and engage the audience through heightened tension.
- Revisit Value Proposition—Answer the value proposition and segue into the vision for the future.
- Call to action—The solution should be very specific and understandable.

Commonalities between the classic Hero's Journey and Linderman's approach to storytelling are twofold, i.e., the goal is to have an objective in mind and communicate a clear message.



Joseph Campbell
and
The Analytic Hero's Journey

STASIS
What is the status quo? How have you benchmarked your environment?

TRIGGER
How did you identify the need for change?

THE QUEST
What are the key contributors and metrics to measure?

SURPRISE
Reality always sets in with unexpected findings?

CRITICAL CHOICE
Can you make a decision based on your metrics?

CLIMAX
The result of the decision.

REVERSAL
Unexpected benefits, consequences, or followups.

RESOLUTION
Current state.

BLUE HILL RESEARCH

© 2014 Blue Hill Research. All Right Reserved

Fig. 2 The Hero's Journey. Eight steps of a story adapted to big data (Park and Haight 2014)

There has been renaissance in storytelling and analysts are approaching many industry segments with the recipes for successful storytelling and have rejuvenated interest in storytelling techniques.

In 2015, Tableau Software presented a white paper, keeping in mind the detective story. The elements of the story—from problem to tension to resolution—appear and are communicated from within the visualization are highlighted [15].

4 Proposed Approach

When data visualization is applied to big data, real-time storytelling is to engage decision-makers. Due to the continuous evolution of social media, complex visualizations require lots of deep analysis and is losing their relevance. Hence, it is critical to develop the accurate story and to reduce the time-bound attention from minutes to seconds.

The vital step in building any real-time self-service data visualization is to develop an easy-to-follow real-time storyboard. Real-time storyboarding is the visual representation of four-part diagram and is shown in Fig. 3 [13].

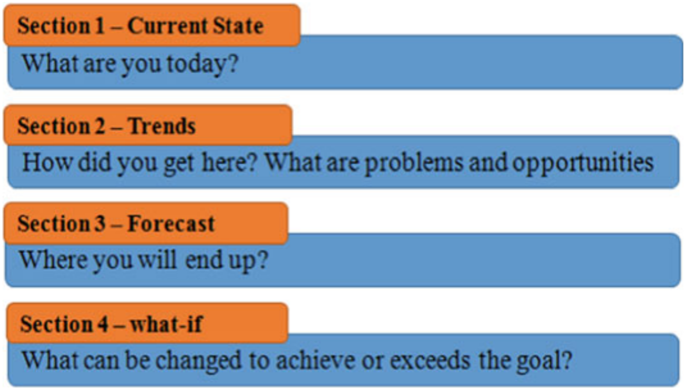

Section 1 – Current State
What are you today?
Section 2 – Trends
How did you get here? What are problems and opportunities
Section 3 – Forecast
Where you will end up?
Section 4 – what-if
What can be changed to achieve or exceeds the goal?

Fig. 3 Storyboard section

Easy-to-follow real-time storyboard is the collection and processing of actionable and particular real-time data that outlines a clear story. To create real-time storyboard, following steps can be followed.

1. Identification of the viewers—Knowing the viewers helps to quickly establish the manner of storyboard to construct and understand the approach for gathering helpful data and visual requirements. If the viewers are C-level executives or senior managers, expect to have little time to inspect immense detail. So data visualization for those viewers must be summarized views that give a past, present, and future of the business with drill down to further required details.
2. Document viewer's goals—A clear and unambiguous understanding of viewer's goals and existing pain points would assist to determine the scope of storyboard.
3. Define KPIs—Understanding the key performance indicator (KPI) that the viewers must view, monitor, or track. It is recommended to keep KPI count to fewer than 10 items combined [13].
4. Dig Deeper to Identify the Sole Purpose of the Story—It is the last step in developing the real-time storyboard. Identify closely on what visualization method is the ideal for better decision-making or analytics. Table 5 is depicting the commonly used data visualization methods. Additionally, identify the KPI which are impactful and are "unknown".

Figure 4 is depicting the KPI fitment chart.

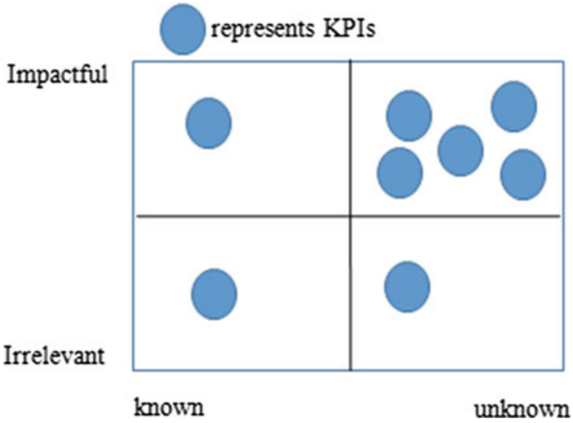
5 Discussion

The art of real-time storytelling is simple and complex at the same time. Stories provoke thought and bring out insights. It is often overlooked in data-driven operations,

Table 5 Big data analysis tool

Sr#	Method name	Big data class
1	Treemap	Applied only to hierarchical data
2	Circle Packing	Applied only to hierarchical data
3	Sunburst	Volume + Velocity
4	Parallel Coordinates	Volume + Velocity + Variety
5	Steamgraph	Volume + Velocity
6	Circular Network Diagram	Volume + Variety
7	Table	Volume + Velocity + Variety
8	Bubble Chart, Scatter Plot, Line Chart, Bar Chart, Pie Chart	Volume + Velocity
9	Word Cloud (Textual data)	Volume
10	Candlestick Chart (Time Series data)	Volume + Velocity + Veracity
11	Map (Geographic data)	Volume + Velocity + Variety

Fig. 4 KPI fitment chart



as it is believed to be a trivial task. What we fail to understand is that the best stories, when not visualized well in real time, end up being useless.

6 Conclusion

In practice, there are a lot of challenges for big data processing and visualization in real time. As all the data is currently visualized by computers, it leads to difficulties in the extraction, followed by its processing and visualization in real time. Those tasks are time-consuming and do not always provide correct or acceptable results.

The main focus of this paper is to give a brief survey of real-time storytelling with self-service visualization techniques that are used in big data. The approach on building any real-time story-telling with self-service data visualization is presented. Hope this paper will serve as a helpful opening to readers interested in self-service visualization and real-time storytelling in big data technologies.

References

1. Chen, C.L.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on big data. Inf. Sci. **275**(10), 314–347 (2014)
2. Jony, A.I.: Applications of real-time big data analytics. Int. J. Comput. Appl. **144**(5) (2016)
3. Mohamed, N., Al-Jaroodi, J.: Real-time big data analytics: applications and challenges. In: International Conference on High Performance Computing and Simulation (HPCS) (2014)
4. Khan, M., Khan, S.S.: Data and information visualization methods and interactive mechanisms: a survey. Int. J. Comput. Appl. **34**(1), 1–14 (2011)
5. Sucharitha, V., Subash, S.R., Prakash, P.: Visualization of big data: its tools and challenges. Int. J. Appl. Eng. Res. **9**(18), 5277–5290 (2014)
6. Data stories—how to combine the power storytelling with effective data visualization, https://www.slideshare.net/miriamgilbert08/data-stories-workshop-34390209
7. HDFS, https://hadoop.apache.org/docs/r1.2.1/hdfs_user_guide.html
8. Hurwitz, J., Nugent, A., Halper, F., Kaufman, M.: Big Data for Dummies. ISBN: 978-1-118-50422-2
9. Spark, https://www.spark.apache.org/
10. Empowering Insight with Tableau: A Case for Self-Service Analytics, https://www.ironsidegroup.com/2016/09/07/tableau-self-service-analytics/
11. Visualizing Big Data: Telling a Better Story—MODSIM World 2018, www.modsimworld.org/papers/2016/Visualizing_Big_Data.pdf
12. Behrens, J.T.: Principles and procedures of exploratory data analysis. Psychol. Methods **2**(2), 131–160 (1997)
13. Data Visualization for Dummies, http://pdf.th7.cn/down/files/1603/DataVisualizationForDummies.pdf
14. 5 Rules for Telling Stories with Your Pitch, https://andrewlinderman.com/2015/08/07/theres-an-easy-fix-for-a-dull-pitch-tell-a-story/
15. The 5 most influential data visualizations of all time, http://www.tableau.com/top-5-most-influential-data-visualizations