



## Simplifying Big Data through Event Chain Manager

Aviral, Sanjeev Thakur

Dept. of Computer Science and Technology, ASET,  
Delhi, India

---

**Abstract**— *Big data is mainly characterized by the large volume, complex, growing datasets with multiple autonomous resources. With the fast development in storage and networking area, the reliability of big data has increased. As now a days, big data is used in many engineering and science domains especially in physical and biological science. This paper presents a concept of viewing the big data as a collection of the events and how to handle the advantage given by this view*

**Keywords**— *Events, Ticks, Event Chain, Temporal System, Event Chain Manager*

---

### I. INTRODUCTION

With the increasing growth of IT equipment, the usage of big data has increased. An event is the smallest unit that happen in the big data entries. It defines a model that is free from the vendor restriction and can be used by the analyst to focus on a specific event chain in the event network of big data. As most of the Business Process Management are interested in the specific data of the big data and perform intelligent analysis. By considering each activity as an event, it gives a more classified view of the flow of information between the events. The relationship between the events can be seen in a more clear view and can be analysed in more efficient way.

The Business Process Management can find the fault and the factors causing it in the event network. Organizations are competing on analytics and the organization that intelligently use the vast amount of data available will survive. So the organization should be free from the constraint of the modelling and using model based system without using the valuable hidden information in the available big data. Every day 2.5 quintillion bytes of data are created and it is up to the organization to pick the valuable information hidden in this enormous amount of information. As big data has a special characteristic that it comes from different heterogeneous sources and with diverse dimensionalities. Different process management requires their own schema to analyse the business operation and extract the meaningful knowledge from the unstructured big data. By employing the event chain in the multidimensional data, compact knowledge of the specific operation can be analysed. This is necessary as it saves time and storage of distributed processing system of the big data. The classified data obtained from this approach is useful for the Business Process Management in this competitive era.

### II. BIG DATA AS A NETWORK OF EVENTS

The event in the event network is analogous to the cell in the human body. It is the smallest entity in the event network and allows the deep analysis of each activity likewise causing factors of an event, event scope, event attributes. The event can be mined from the internet or some usage of physical devices enabled by the IT e.g. ATM. So events can be categorized as:

Event of People (EOP): the data related to social interaction.

Event of Things (EOT): the data related to usage of physical objects likewise usage of ATM card.

Event of Content (EOC): the data related to the information send from various sources likewise web pages, YouTube, newsfeeds, Wikipedia etc.

Event of locations (EOL): the data related to the geographic location likewise the latitude and longitude information expressed by the GEOLASH. The GEOLASH was invented by the Gustavo Niemeyer in 2008. It express the geographic location as a matrix of latitude and longitude by simply using the URL [www.geohash.org](http://www.geohash.org)

These events partially overlap and can have relation to each other. This partial overlap property helps in the information flow in the event system. This property of event system helps in tracking a specific event chain and analysing the properties of that specific event chain likewise:

To detect the blockage in the event information flow and analysing the factors causing it.

To detect the ambiguity and understand the severity of that ambiguous event

To predict cost, risk and delays

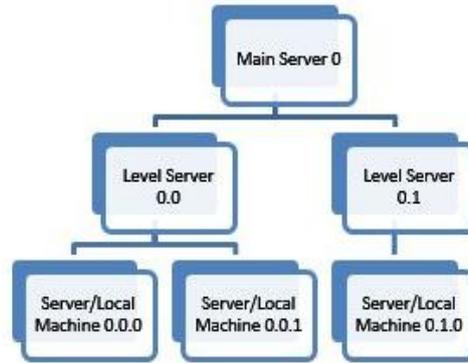
To backtrack the faulty source event in the event chain

To recommend actions and avoid inefficiencies.

These four event types are not in form of directly analysed as big data is itself complex. So these event are first to be extracted from the event store. After that they are refined, filtered and converted into an event chain. While all these events types require a standard timestamp server system which is fulfilled by the temporal server system.

**Temporal Server System:** A hierarchy of temporal servers have to maintain to provide timestamp to all the devices so that they will be uniformly synchronized and can be standardized. The timestamp is allotted to the event at both entry and the exit time through which the duration of the event can be detected easily. The network of temporal servers is responsible for the uniform order of the events in a timely fashion and to handle the ambiguity for the events that happened at the same moment of the time.

In this hierarchal system, one server passes a special key called Tick for the identification of the specific server. The main concern is to form a chain of these Tick's along with the previous master servers. [9]These Tick's follows the Temporal Relation which gives the Temporal Pattern in the form of chains. This Tick chain helps in the locating the source event of a specific event chain and helps in analysis of the event system likewise to detect the bottleneck condition the path between different servers are analyzed. Mutual exclusivity of the Tick is assured by the mining algorithm consists of candidate generation phase and counting phase. Every Tick is unique and has columns to identify the sender and recipient server to maintain uniqueness. This approach offers uniqueness and complete the expectation of hierarchy in more efficient way as they can be backtracked easily by following the series of addresses and events can be related in more efficient way



Proposed Temporal server architecture

As discussed in the [1], [2] and [9] the time dependent data follows the association rule and forms the series of Ticks which helps in backtracking of the flow of information and to detect the defect in the chain.

As discussed in the [3] and [4], Ticks formation from the information provide a flexible view to gain insight information.

**Issues in temporal system:**

**Delay due to geographic distances:** As due to various geographical conditions likewise mountains, deserts the signal propagation may be suffered. To counterfeit the some specially equipped devices should be used. The satellite system is also favourable in that situation if that occurs within the budget.

**Fault tolerance:** As due to vast geographic conditions, the connectivity may suffer.

**Processing time of various devices:** There is a need of synchronized devices so that there will be no time delay to backtrack the events in the big data. Efficient fast processor is the main requirement in the devices that allows parallel execution in fast way.

Parallel computation in the server is avoided.

**Temporal Data:** Intuitively, all the events in the event network of big data has temporal characteristic, so they can be classified in the following way:

**Entry temporal data:** the data which is obtained by the first encounter with such a defined interface such as face book login, ATM card put in the machine etc. This data is mainly obtained by the temporal server system. It is essential as it removes the ambiguity (equipped with operational data).

**Operational data:** the data which describes the operation done in the time slot. This is the data that discriminates the objective from others. By giving each objective likewise face book operation and call a different event type. It becomes easy to prioritize the specific event in the event chain. These event types belong to four different types of events (EOP, EOT, EOC, and EOL).

Table 1

Event Number	Tick	Event Entry Timestamp	Event Operation Data	Event Exit Timestamp
1	0.0	20-6-2015 18:10:02	ATM	20-6-2015 18:11:03
2	0.0	20-6-2015 18:12:06	e-mail	20-6-2015 18:12:20
3	0.0.1	20-6-2015 20:20:04	e-mail	20-6-2015 20:20:18
4	0.0.0	20-6-2015 22:10:10	Call	20-6-2015 22:15:10
5	0.1.0	20-6-2015 22:10:10	Call	20-6-2015 22:15:10
....	....	....	....	....

Exit Temporal data: this data is also obtained by the temporal server system to symbolize the operation data ending. They play a major role in linking the various operational data in an objective to give a backtrack approach of the events. Among these, the entry temporal data and exit temporal data serves as an events checker and gives an indication of duration of an event. As an individual operation data is of no value to come to a conclusion for a specific event chain so all the information flow between the events is analysed as below:

### III. EVENT CHAIN PROCESSING

The ambiguity between the events can be resolved by the TICK associated with the events entry data and events exit data. As per the [5], various issues are encountered in processing the big data which can be solved by the event chain. As in table 1, the event number 4 and event number 5 are differentiated by the pairing of the TICK with event entry data and event exit data. And same approach is used for distinguishing the event 1 and event 2 which has same TICK as they have different event entry data and event exit data.

As in event chain, events overlap each other and passes information to each other. So a data flow diagram can be easily be drawn for the event chain. The most interesting aspect of the event chain that it is unique for the same length of the event network of big data where it gains the idea of mutual exclusivity if Ticks as per the [9]. In contrast to [6] and [7], this event chain model provide cross validation and hence flexibility for the dynamic time dependent data. As per the [8], this model gains the idea of spatio-temporal indexing in the form of the Ticks and its hierarchy model.

Length of event chain: Numbers of Tick involved in the event chain. As the number of ticks increases the complexity increases linearly. But it is not the main cause of the increasing complexity of the event chain as complexity can occur due to demand of resource by the event during the processing time.

Event forest: It is a network that originates from one event and proceed in unguided manner to many events. It is the responsibility of the Event Chain Manager to select the path from the forest. Event forest is analogous to simple big data.

As in table 1, the different operation performed at different timeslot is mentioned and the relationship between these events are analyzed by using the Tick combined with the time stamp.

The Event Chain for the Table is as in Figure 1:

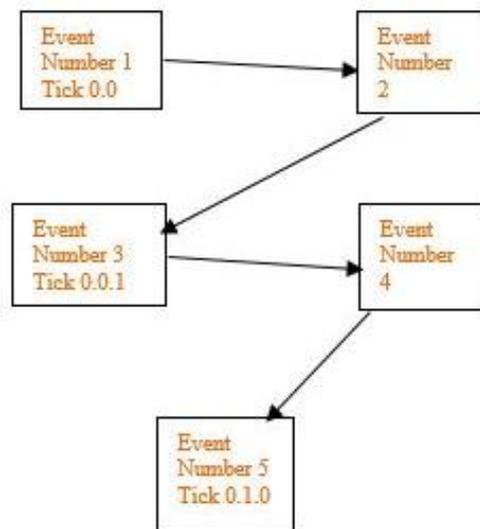


Figure 1

Event Rate is the events happening in the event network for a specified event chain. This plays an important role for the election of the event chain in the event chain manager. It is basically the efficiency of the event chain manager and much more depend on the status of the event chain.

**Status of Event Chain:** It is not always the events happening in the network are useful to the specific event chain and there can be situation that the rate of events in the network takes a lot of time. In that situation event chain processing becomes a burden on the event chain manager. To avoid these conditions, different status of event chain are maintained.

**Active status:** It is the status of the event chain when the event rate in the network is specifically approachable to a current event chain in the event chain manager. It saves the pre filtering time and gives a direct flow of events to the current event chain in the event chain manager. For example: events happening in the earth rotation and different solar radiations are continuous events. So that event chain is always active in specific event chain manager of climatic analysis program.

**Hibernate status:** It is the status of the event chain when there is a guarantee of the event happening in the future but that takes a little more time. This situation can be happened in the case of dependency of event on the shared resources. By applying this status, the event chain can be put to hibernate and event chain manager can be employed for another event chain during that little more time. For example: the events of the important web page access can be put in hibernate mode as there is a guarantee of event rate but not in continuous flow.

**Dead status:** It is the status of the event chain when it is not required or the information has been extracted from that event chain. It is useful as it saves the storage space and differentiate the interest area for a specific operation. For

example: event chain of a person medical report can be put to DEAD state after the person is no more and there is no upcoming event probability.

Suspend status: It is the special case of Hibernate status as it convey the error in the event growth in the event chain .When the event chain manager identify the error in the event chain, it puts the event chain to SUSPEND status and saves the object of that event chain .By using this approach the backtracking the events in the event chain is performed by using object. The site of error is identified and event chain new object is saved for the just previous sub-event chain in the event chain Until the event chain get a new object the event chain has SUSPEND status.

#### IV. EVENT CHAIN MANAGER:

The main task of the event chain manager is make an object for each individual event chain and characterize the event chain by that. Only the authorized event are allowed for further processing in the event manager. It performs the bit-by-bit growth for each event addition in the specific event chain with the use of sequencer. Sequencer performs the addition of events in the event chain whose object has been created. It has the responsibility to authenticate the event by the TICK such as in the above table 1. As discussed in the [11], [12], [13] and [14], it also works as a tracker and filter for the incoming event in the analyzing system likewise of the Business Process Management. It handles the orientation of the analysing system in the event forest. The different algorithms for the decision making and programs are employed in the event manager .It is basically the core of this whole approach.

Event chain manager performs the processing of the event chain by using the status information of the event chain stored in the event chain object. The individual status of the event in the event chain is focused that makes it the flexible .It is done with the use of event chain status checker. As each event from the event chain is processed individually from the event chain, so at that particular tick, the status of the event to be processed becomes the status of the event chain. By using the object oriented approach, it becomes easy to track the event chain status and also to identify the status of the event chain. As in SUSPEND status, until the new event chain object is made the event chain is in SUSPEND status. As per the [15] and [16], Event chain manager can use the deadlock detection algorithm in case of deadlock between the events. As big data is distributed in nature so a distributed deadlock detection algorithm Transaction Wait for Graph combined with probe generation mechanism is used here. Each event of event chain is dependent on the previous one i.e. Tick (i) ->Tick (i+1), this technique allows multiple parallel execution of event chain and resolve the priority dispute by probe generation. By using this,

Certain event chains from the event forest can be processed simultaneously.

The clustering of events require the communication between them to form an event chain. To understand their common behaviour towards the aim as it is achieved by using the tick and operational data inside them.

In Figure 2, the various component of the event chain manager are depicted in their role order for a specific event chain.

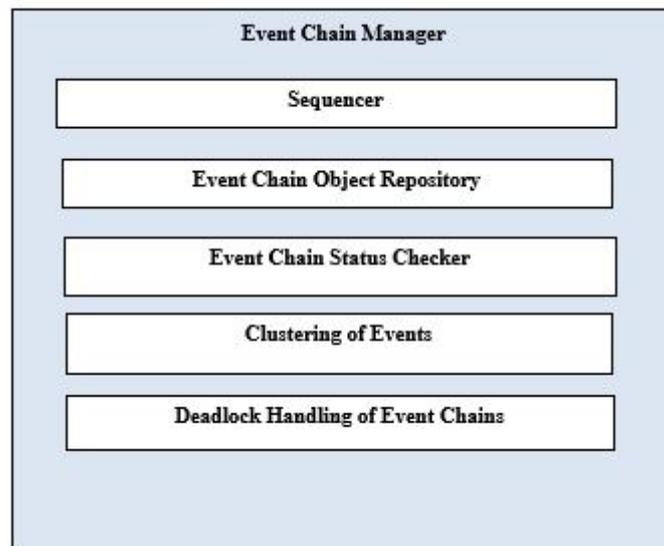


Figure 2

**Communication:** As each event chain is a group of events that are logically or relationally linked to perform the flow of information.

In other words, each event chain can be considered as a cluster of events. As in table 1 each the flow of information is characterized by the associated Tick. Several information is analysed from the TICK.

Tick (i,j,k)

i->Geographical id

j->Server id

k->Remote Machine id

If

k is incremented only

Operation is performed on single remote machine i.e. ATM

If

{

j is incremented, server location change is indicated and hence the movement of the target is considered.

If

I is incremented, geographical condition change is indicated

By the various combination of these (i,j,k), the target is tracked and hence the event chain is built. And each tick has its temporal data through which the events can be arranged in increasing order. The increasing order of related ticks communicate by message passing of its temporal data to the sequencer in the event chain manager. The sequencing of the related ticks in the increasing order to form an event chain is performed by the sequencer of the event chain.

**Clustering of Events:** Clustering is the fundamental step of data analysis. It is the task of assigning a set of objects to group. As each event of event chain is characterized by the basic four types of events which distinguish it with a combination of Tick from other events of event chain. Each event chain is considered as a cluster of events related together by specific relation. As in an organization, the events occurred in specific department form a separate event chains. As per the [17], many different clustering validity measures exist that are helpful in quantitative measure for evaluating the data partition performance to characterize the similarity between the of employees events, the relativity between the events is used. As per the [18], the relative clustering perform the cross validation between the event chains to perform the clustering. It is mainly due to the adopted reference partition of the data.

In contrast to [19], the reference matching in the clustering can be avoided as it solely depends on the individual Ticks and their relationship. Tick plays its role in indexing the events as in table 1 and as shown in the event chain of table 1.

## V. EXPERIMENTAL RESULTS

For the experiment a data of 200 employees of an organization is collected, the activities form a separate event chains for a specific employee. Each event in the event chain is also associated with the Tick for the temporal analysis. So for calculation of the faulty employee in various department who does not follow the policy of the organization, the event chains are processed.

The results are depicted in the figure 3:

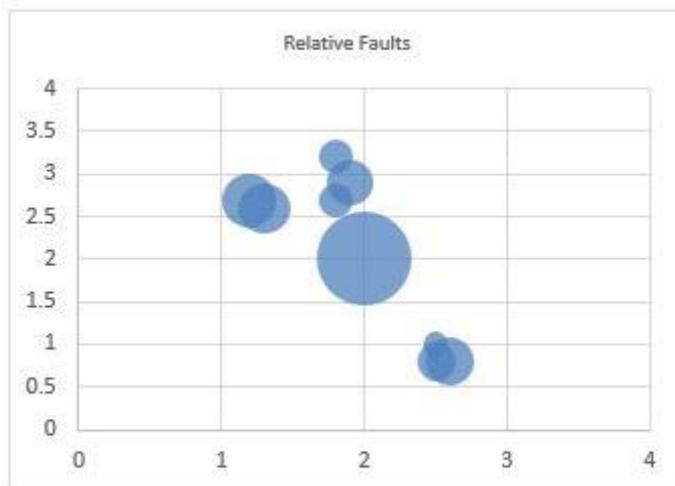


Figure 3

The X-coordinate express the Tick and Y-coordinate express the relative faults done by the employee. So it can be analysed the impact of fault done by the employee is highest at the tick of 0.0.2.

## VI. CONCLUSIONS

As per the [20] and [22], the clustering of events generated by the activity of employees are categorized into four main classes of events and the clustering of events are performed dynamically and with flexibility. It provides the analyst to define the relation according to the need dynamically to perform the K-means and Hierarchical clustering. By following the event approach in big data, the tiny section of a big data can be focused and analysed. Flexibility and dynamic decision making are the main profits of this approach. Event based model provides the freedom the problems as discussed in the [21], as each event chain manager follows the status of the event chains and they can be modelled similarly to the threads as in Java.

### Benefits:

Freedom from Relationship Complexity: It is necessary as volume of big data is increasing day by day, so do the complexity and the relationships underlying the data. As an event occurrence is independent of the other events residing

in the other event chain, hence a partition is made between the interest area and the rest. This type of bit by bit growth of event chain makes the event and its properties more open to the analysing system. Hence a single event does not carry valuable information about the event chain but the properties of individual event can be modified according to the requirement. This freedom from the relationship complexity gives the dynamic support to analyse the system, which is the main requirement in today's competitive business process management. So it performs:

**Abstraction:** As only selected event chain's event are analysed it gives much more simplified view for the Business Process Management.

**Encapsulation:** It also performs the encapsulation of the portion of the data from the rest big, unstructured volume of Big Data.

**Processing and Network favourable:** As Big data operation is distributed in nature but still it requires high performance CPU. By this event approach on the Big Data, only the N sub-section of the big data is transported and processed which will increase the performance of the cut and slice operation and Map Reduce algorithm of the big data.

**Freedom from semantics:** Each event is characterized by its event type so the semantic constraint is reduced to very much extent.

**Reduction of Dimensionality:** Each event is characterized by its event type and it conveys about the information of only its own state. By selecting only the interested information from the event it becomes easy to reduce the dimensionality. By integrating this approach to the whole event chain system, it becomes easy to reduce the dimension of the abstracted data from the Big Data.

### ACKNOWLEDGMENT

I would like to express my deepest gratitude to my guide Mr Sanjeev Thakur for his excellent guidance, patience and providing me the excellent atmosphere for doing research. This research requires further advancement for its implementation and provides a great scope for its future uses.

### REFERENCES

- [1] R. Agrawal and R. Srikant "Fast Algorithms for mining association rules in large databases". In Proceedings of VLDB, 1994.
- [2] P. shan Kam and A. W. chee Fu. "Discovering Temporal patterns for interval based events" In Proceedings of DaWaK, 2000.
- [3] F. Hoppner "Knowledge Discovery from Sequential Data" PhD Thesis Technical University, Braunschweig, Germany 2003
- [4] P. Papadimitriou, G. Kollios and S. Sclaroff "Discovering frequent arrangements of temporal intervals" In Proceedings of ICDM, 2005.
- [5] Laurance T. Yang, Jinjun Chan "Special Issue on Scalable Computing for Big Data" Science Direct, Big Data Research Volume -1(2014), pages 1-66.
- [6] Craig C. Douglas "An Open Framework for Dynamic Big Data Driven Application Systems (DBDDAS) Development" In Proceedings of International Conference on Computational Science, Science Direct, pages 1246-1255.
- [7] Jianjun Yu, Fuchun Jiang, Tongyu Zhu "RTIC-C: A Big Data System for Massive Traffic Information Mining" Published in Cloud Computing and Big Data (Cloud-Com Asia), IEEE 2013, pages 395-402.
- [8] Anthony Fox, Chris Eichelberger, James Hughes, Skylar Lyon "Spatio-Temporal Indexing in Non Relational Distributed Databases" In Proceedings of Big Data, IEEE International Conference, 2013, pages 291-299.
- [9] Iyad Batal, Dmitriy Fradkin, James Harrison, Fabian Moerchen, Milos Hauskrecht "Mining Recent Temporal Patterns Event Detection in Multivariate Time Series Data" In proceedings of ACM, 2012.
- [10] Tian Guo, Thanasis G. Papaioannou, Kerl Aberer "Efficient Indexing and Query Processing of Model View Sensor Data in the Cloud" Science Direct, Big Data Research 2014.
- [11] Carlo Zaniolo "Event Oriented Data Models and Temporal Queries in Transaction-Time Databases" Published in Temporal Representation and Reasoning, 2009, IEEE, pages 47-53
- [12] Hassan Sayyadi "Event Detection and Tracking in Social Streams." Proceedings to 3<sup>rd</sup> International Conference of AAAI on Weblogs and Social Media, 2009.
- [13] Badrish Chandramouli, Jonathan Goldstein, Songyun Duan "Temporal Analytics on Big Data for Web" In Proceedings of 28<sup>th</sup> International Conference on Data Engineering, ACM, pages 90-101
- [14] Junyu Xuan, Xiangfeng Luo, Jie Lu "Mining Websites Preferences on Web Events in Big Data Environment". Published in Computer Science and Engineering (CSE), 2013 IEEE 16<sup>th</sup> International Conference in December, pages 1043-1050.
- [15] Xiaoyong Li, Yijie Wang, Xiaoling Li, Xiaowei Wang, Jie Yu "An Efficient Approach for Skyline Queries over Distributed Uncertain Data" Available at Science Direct, Big Data Research.
- [16] Swati Gupta "Approaches of Deadlock Detection and Prevention in Distributed Systems" International Journal of Software and Web Science 9(2), 2014, pages 86-88.
- [17] M. Halkidi, Y. Batistakis and M. Vazirgannis "On Clustering Validation Techniques" Intelligent Information Systems Journal, Kluwer Publishers, pages 107-145, 2001.

- [18] Lucas Vendramin, Ricard J.G.B. Campello and Eduardo R. Hruschka” *Relative Clustering Validity Criteria: A Comparative Overview*”.
- [19] Andrew McCallum, Kamal Nigam, Lyle H. Ungar”*Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching* “In proceedings of the sixth ACM SIGKDD International Conference on Knowledge discovery and Data mining ,169-178
- [20] Amandeep Kaur Mann, Navneet Kaur “*Survey Paper on Clustering Techniques* “Published in International Journal of Science, Engineering and Technology Research(IJSETR) Volume-2,Issue-4, April 2013.
- [21] Nitin Agarwal, Ehtesham Haque, Huan Liu and Lance Parsons “*Research Paper Recommender Systems: A Subspace Clustering Approach* “In Proceedings of 6<sup>th</sup> International Conference on Advances in Web-Age Information Management, pages 475-491.
- [22] Aastha Joshi and Rajneet Kaur “*A Review: Comparative Study of Various Clustering Techniques in Data Mining* “Published in International Journal of Advanced Research in Computer Science and Software Engineering Volume 3.