

# Data analytics for Cyber security

- Big Data Analytics  
and Its Need for  
Cybersecurity-

Vandana P. Janeja

©2022 Janeja. All rights reserved.



# Outline

What Is Big Data?

Big Data in Cybersecurity

Landscape of Big Data Technologies

Mahout and Spark Comparative Analysis

Complex Nature of Data

- Nature of Data: Spatial Data
- Nature of Data: Graph Data
- Nature of Data: Other Types of Complex Data

Where Does Analytics Fit in for Cybersecurity?

Change Detection in Massive Traffic Datasets

Multipronged Attacks

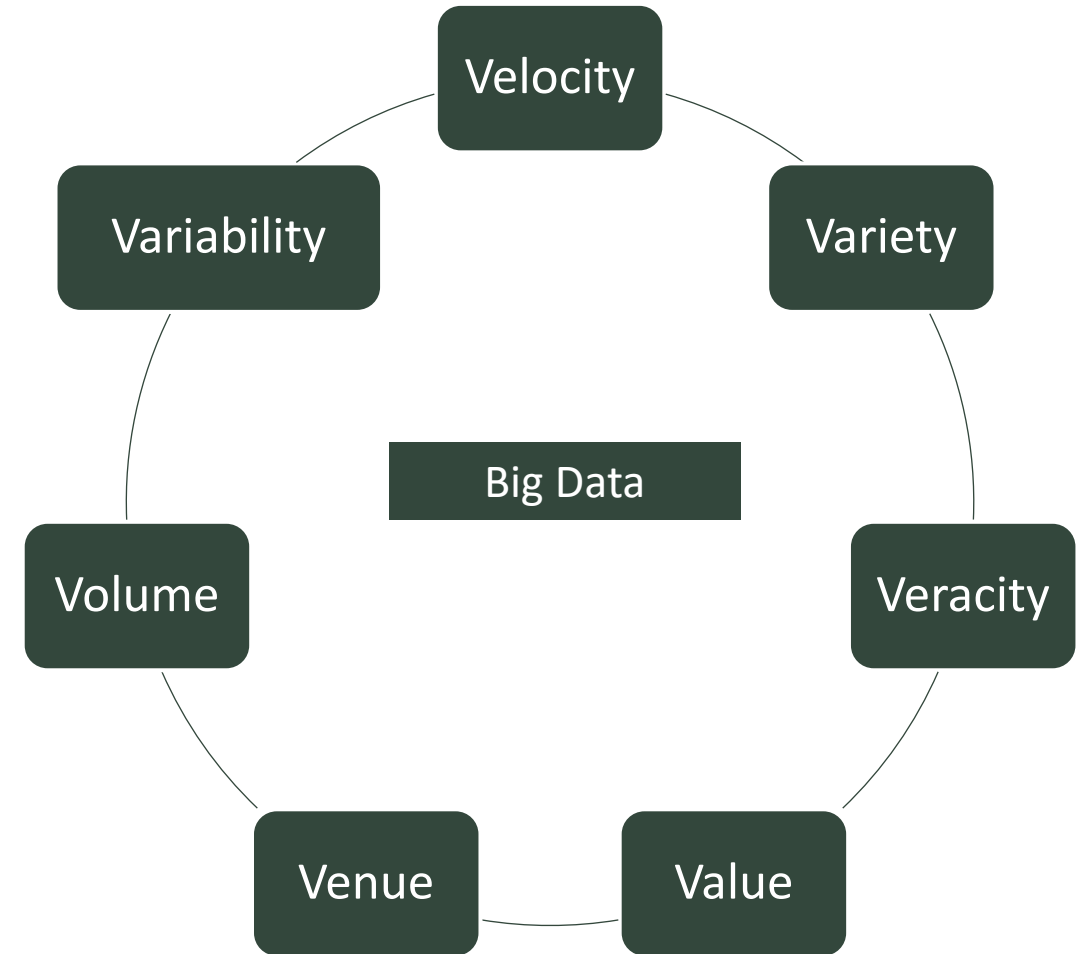
Privacy and Security Issues in Big Data

# What Is Big Data?

- The nature of big data comes from the complexity of the data and the mechanisms required to bus the data, analyze it and find insights from it.
- Several initiatives have been established to define what we mean by big data (such as NIST BDWG (NBDIF 2015), Ward& Barker (2013), Bayer & Laney (2012)). They highlight these parameters as the V's of big data. There are 3-7 or more of these V's based on the different types of definitions, as shown in the figure 4.1
- In some cases data which may not be particularly large but has many of these other qualities such as variability, variety, and value may also be considered big data.
- The term Big may refer to the big aspect of the overall complexity of the data. If a dataset is highly unusual in terms of its complexity it may qualify as big data under one of the big data qualities.
- Cybersecurity domain
  - Pure network data traffic - it is massive, for example in a mid-size organization it can range into petabytes per second.
  - Internet of things - the complexity of data is based on the velocity and variety and in some cases also the volume. One example is a sensor network which generates a constant stream of data originating from multiple sensors over time producing a high variability in the data due to the environmental factors.
- The value from these big datasets can be derived by sifting through these complex datasets to derive insightful patterns.

# Big Data - V

- Big data does not just refer to the Volume or size of the data
- Data can be complex and have the following qualities:
  - velocity- generated at a rapid pace,
  - variety-consist of multiple types of heterogeneous data,
  - veracity-provide trustworthy insights into the domain function,
  - value-may be able to generate revenue or provide other benefit,
  - venue-be dynamic with respect to the location, volume-large amounts being generated,
  - variability-some aspect of these changing over time.



# Big Data in Cybersecurity

# Internet Users

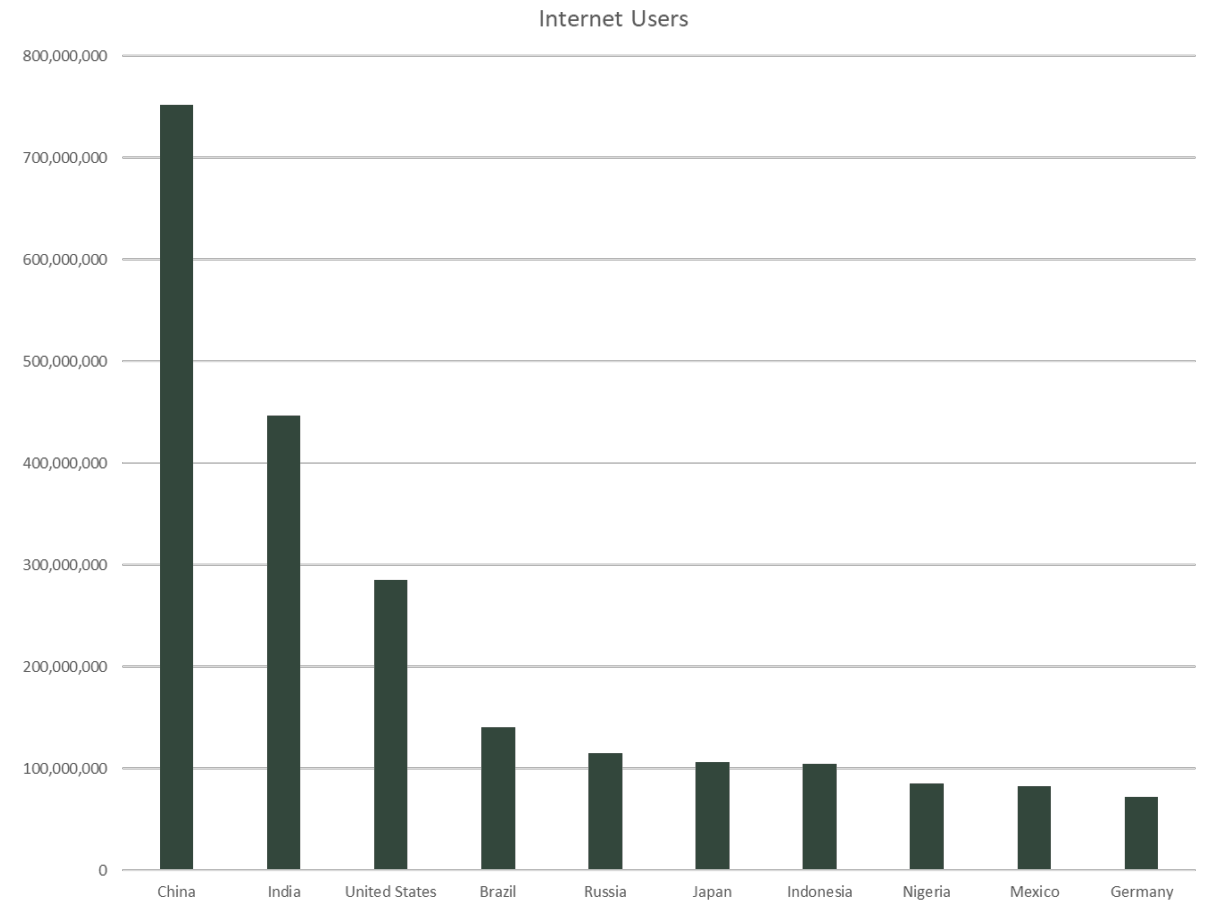
According to the world fact book maintained by the CIA (Factbook 2017) the number of internet users is now in the billions.

The top 25 countries in terms of the internet users.

Vast number of users on the internet are generating traffic from computers, mobile devices, things/other devices.

The network traffic data is truly complex in terms of the variety of the data generated, volume of the data and velocity by which it is generated.

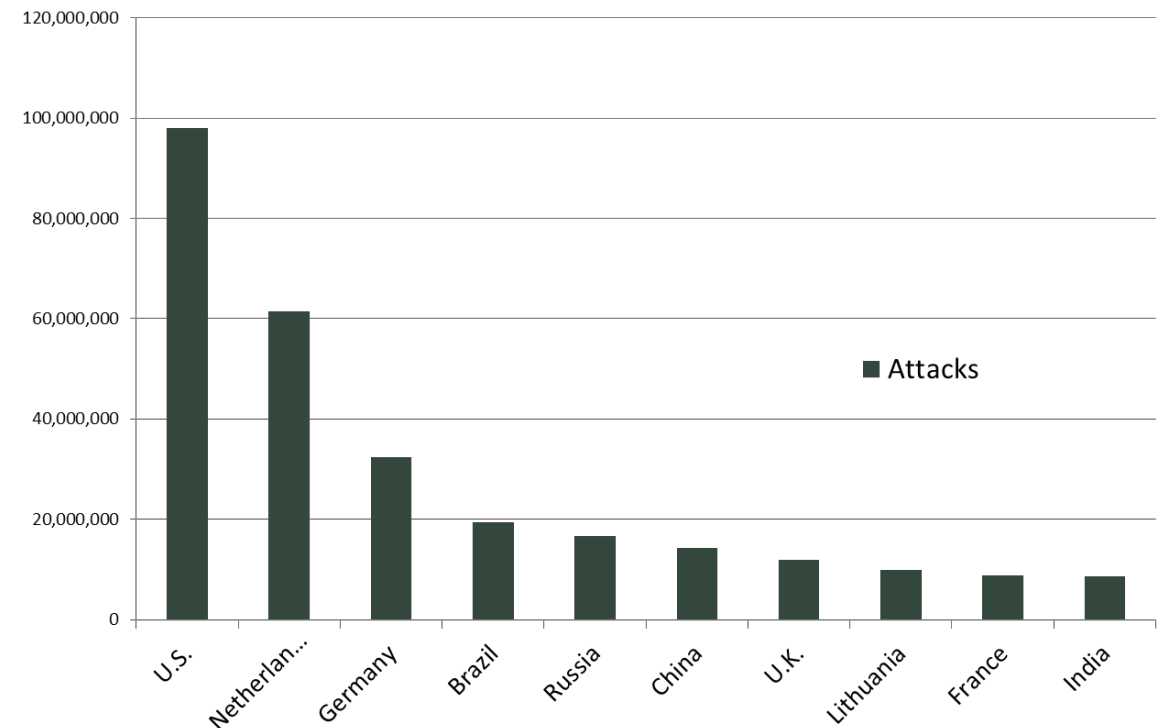
There are also organizations which are generating the data and network traffic.





# Attacks Sourced from Countries

- While majority of the attackers may come from countries with majority internet users, in addition, they may also come from the countries with fewer internet users.
  - For example, number of internet users in Lithuania is roughly 1/3rd of the United States and is 97th in the list of ranked countries by number of internet users. Despite of this Lithuania is still placed 8th in terms of the countries sourcing cyber-attacks.
- Clearly the amount of internet traffic generated is not an indicator of attack traffic.
- There are other factors at play in identifying cyber-attacks on the larger world stage taking into account not only the internet traffic but socio-political factors, crime rates, other cyber-attacks.
  - For example, a simple search of Lithuania and cyber-attacks reveals that Lithuania is also at the receiving end of many major cyber-attacks. More recently, Lithuania led a cyber shield exercise to practice procedures for protecting the cyber infrastructure (L24, 2016). Thus, it is not clear whether the sourced attacks, where Lithuania is ranked 8 across countries in the 4th quarter of 2016, are more defensive or offensive.
- It is clear that a deeper dive can reveal a lot more insights into the real state of affairs on the internet.
- If we consider an organization and study trends of usage and simple rules to analyze traffic we may miss insights of potential advanced persistent threats.
- Such threats can be evaluated by not just looking at data in isolation but in combination with user behaviors, socio-political factors, context of time and potentially many other contextual features.
- This is truly a big data problem as it brings in the volume, velocity, variety, variability, and venue and can lead to great value with high veracity of the results. Here testing the veracity is equally important to generating value.
- The challenge is to sift through the internet traffic and identify attacks. This can be at the device level, user level or the network level at large.



# Landscape of Big Data Technologies

Big Data technologies can be selected based on criteria that fit a business need

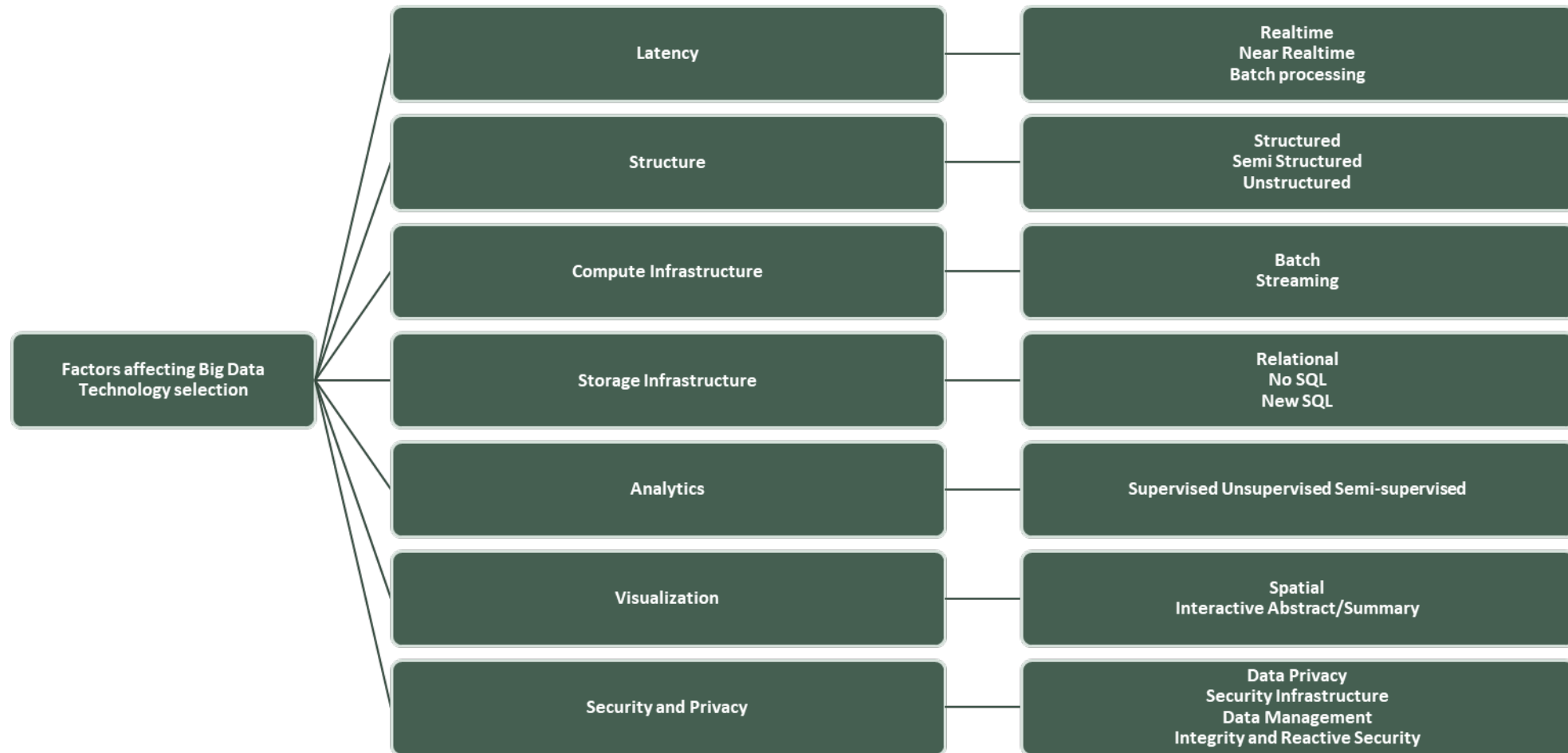
- latency in response time,
- whether the data is structured, if a SQL like environment is required,
- types of analytics required,
- specific types of visualizations that may be needed and
- the security and privacy needs of the business.

Big Data technologies can be selected based on these criteria.

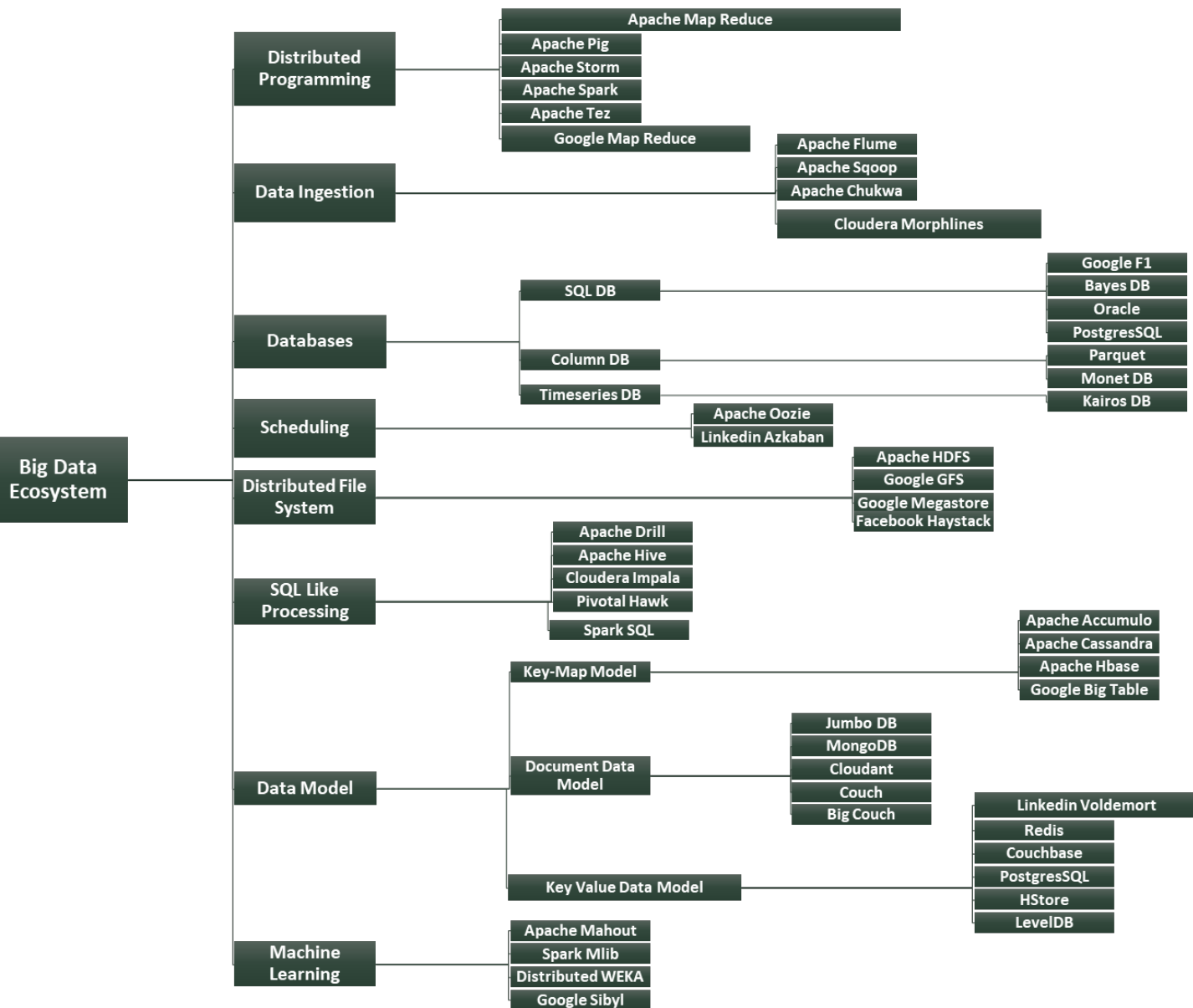
- For example if the need of the business is low latency and SQL like processing then those tools can be selected which have a quick turnaround time for queries in massive datasets.
- Of the several big data analytics frameworks present in the market, the business can select the tools that provide Massively Parallel Processing (MPP) such as engines on top of Hadoop that have high SQL like functionality and portability such as Apache Hive, Facebook Presto, Apache Drill, Apache Spark.
- Out of these Presto and Spark have been shown to produce better outcomes for SQL like processing.



# Big Data Technology Selection-Factors



# Big Data Technology Landscape



# Complex Nature of Data

The premise of achieving a high level of analysis and mining in big data, is a good understanding of the nature and complexity of the data

Mining in large heterogeneous data is increasingly becoming challenging due to the complexity of the data

Many data mining algorithms have been proposed for small datasets

The nature of the data itself renders a challenge to the data mining

Mainly such problems can be exemplified in the mining of high dimensional data sets

The quality of the mining results is affected adversely as the number of dimensions increases. Here dimension refers to the attributes of an object

Outlier detection i.e. the identification of anomalous objects

- In higher dimensions data becomes sparse, points tend to become equidistant.
- This adversely affects the methods based on density of points, unless data follows certain simple distributions.
- The outliers could be hidden in high dimensions.
- For outlier detection a subset of the entire attribute set could be used to detect or indicate the outliers.

# Nature of Data: Spatial Data

- Spatial data mining deals with identification of non-trivial and useful knowledge discovery in spatial database where spatial (point, lines, polygons, location, pixel data) and non-spatial data, e.g., population count are stored
- Unlike traditional data mining it is important to address the nature of spatial data, which renders new challenges in the inherent spatial autocorrelation and heterogeneity in the data. Spatial data is seen in the cybersecurity domain in sensor networks placed in a region
  - These are particularly relevant to cyber physical systems which have interactions with several physical sensors which are impacted by the regional variables and also non spatial elements such as the computer networks involved in the communication and analysis of the data.
  - CPS provides a very comprehensive example of how the cyber elements interact with the physical elements such that to perform any type of knowledge discovery the physical elements along with the environmental variables impacting the physical locations need to be modelled correctly.
- Spatial data poses more of a challenge due to the type of data, volume of data and correlation between the objects and their neighbor or neighborhood
- There can be relationships between spatial objects like topological relationships, direction relationships, metric relationships or complex relationships, which are based on these basic relationships.
- The data could depict temporal changes, which is inherent in the data itself.
- The same spatial data can also be represented in different ways i.e., raster and vector format. For example, georeferenced data, includes spatial location in terms of latitude, longitude and other information about spatial and non-spatial attributes
- Spatial data can benefit cyber security applications in characterization of the region with socio-economic data and cybercrime data
- This characterization can help determine which cities, areas, or countries may be more likely to send a malicious request
- Spatial data can be utilized to produce a cyber-security map which can combine several data streams and provides a geospatial interface to visualize the location of cyber activity

# Nature of Data: Graph Data

Graphs represent a set of nodes and links between the nodes

With the proliferation of computer network data graphs are becoming massive in size and are constantly evolving

- Example 1: Nodes can be IP addresses and the links can be the packets or communication sent between the two IP addresses.
- Example 2: Router network where the nodes are routers and links are the possible paths from the router table.

This is an example where big data techniques can facilitate the discovery of novel insights

- For example, how does a graph for a computer network traffic change over time in terms of node level and graph level properties can help determine potential events
- If a node is highly connected it is an important node
- We can study the behavior of the network data whether it increases the diameter of a graph or density of the graph
- These properties have direct implications on managing the networks, identifying critical events such as cyber-attacks and evaluating the impact of events on the graphs

# Where Does Analytics Fit in for Cybersecurity?

Traditional cybersecurity methods looking at network traffic data focus on Intrusion detection which aims at identifying threats of unauthorized access using signatures of unauthorized access or attacks


Data Analytics can go beyond this signature based discovery to identify complex types of attacks which may be hidden in massive datasets or may be spread out over time and multiple networks

Some examples which show the need of using big data analytics in cybersecurity:

- Sampling based change discovery in massive networks (Namayanja ET. Al 2015)
- Big Distributed Intrusion detection system (Janeja ET. Al. 2014)

# Change Detection in Massive Traffic Datasets

Computer networks can be seen as graphs of communications between nodes, where each node is represented by an IP address



Consider a snapshot of network traffic data - generate a communication graph from it



We can evaluate whether there have been node level or network level changes in the graphs indicating potential cyber-attacks

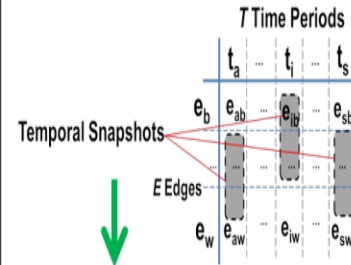


# Change Detection-Example

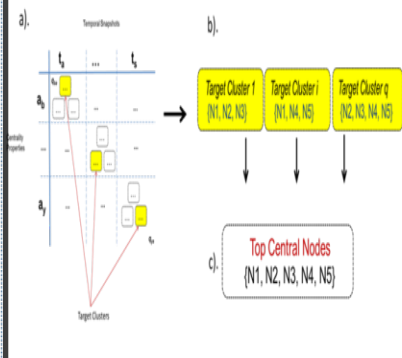
- Given that network communication can be very massive it is important to identify subset of this traffic to evaluate
- This can be done by targeted sampling, one way is to select the top central nodes where nodes with very high degree of connectivity represent the major communication in the traffic data
  - If these central nodes are consistent through time, represented by bins or intervals of time, then we can say that these nodes are consistent.
  - If these nodes are not present in all time periods then they are inconsistent nodes.
  - If there are time periods where several of the central nodes become inconsistent then this time period is a potential time point where an event affected these nodes and may require further investigation
  - The sub graphs for the time intervals can also be studied to evaluate graph level properties
  - If the densification or diameter does not follow the network properties during time periods it can again be an indicator of a network level event at the time points

## Node Selection

### 1. Random Temporal Network Snapshot



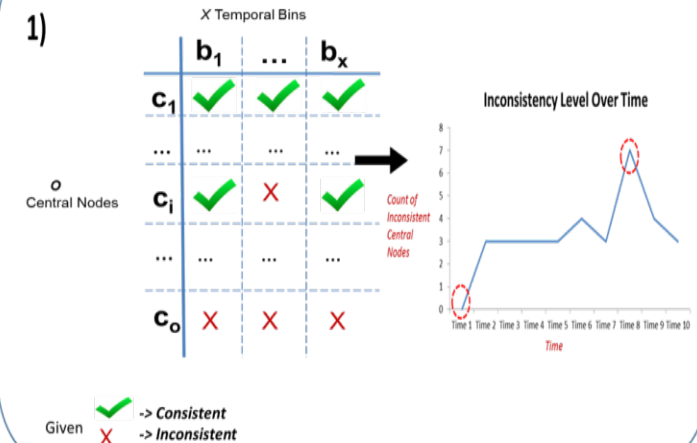
### 2. Target Cluster based Centrality Sampling



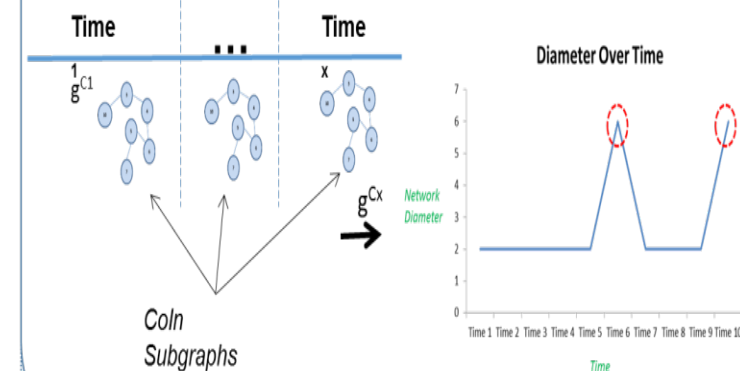
## Multi-Level Change Detection

### Time Periods of Change due to Coln Central Nodes (Coln-TPC)

1)



### c. Network Level Coln Change Point Detection



# Change Detection- Example

---

- Let us consider the graphs across three time periods where the degrees of the nodes
- Each graph represents a set of communication lines between the IP nodes
- Note that here the degree is based on the non-duplicate edges
- This can be modified to a weighted degree to count for multiple times the nodes communicate with each other, for accommodating duplicate edges
- Alternatively edge weight can be added for the number of times the nodes are communicating
- For simplicity we only consider the non-duplicate communications
- We can see that certain nodes (a, d, e) are consistently central if the degree threshold for centrality is greater than or equal to two. On the other hand we see that node 'f' is consistently low degree
- Node 'b' starts as a high degree but is dropped in time period T2 and again comes back as a low node degree in T3
- The consistent nodes a, d, e can be seen as the regularly used nodes if they are consistently central
- If such consistent nodes get dropped after being consistent in several time periods then this can prompt further analysis
- If several nodes lose their degree in a time period then we can also label that time period to prompt further evaluation
- This process helps identify central nodes over time to evaluate any unusual changes in the communications

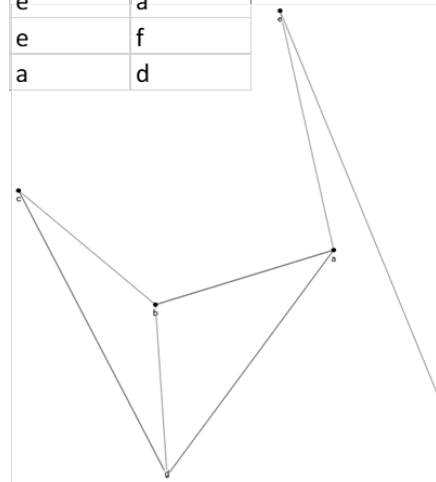
# Change Detection-Example

IP1	IP2
a	b
a	b
a	b
a	b
a	b
a	b
a	b
a	b
a	b
a	b
c	d
c	d
c	d
c	d
a	d
a	d
b	d
b	c
e	a
e	f
a	d

T1

Nodes	Degree T1
a	3
b	3
c	2
d	3
e	2
f	1

Central Nodes a,b,c,d,e

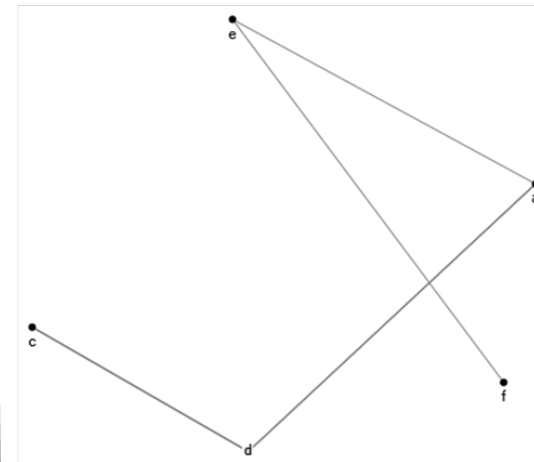


IP1	IP2
c	d
c	d
c	d
c	d
a	d
a	d
e	a
e	f
a	d

T2

Nodes	Degree T2
a	2
c	1
d	2
e	2
f	1

Central Nodes a,d,e

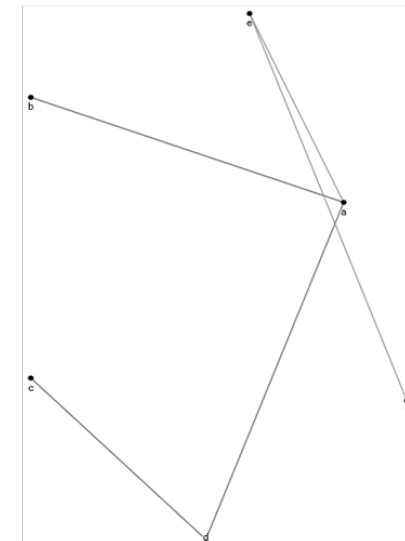


IP1	IP2
a	b
a	b
a	b
a	b
c	d
c	d
c	d
c	d
c	d
a	d
a	d
e	a
e	f
a	d

T3

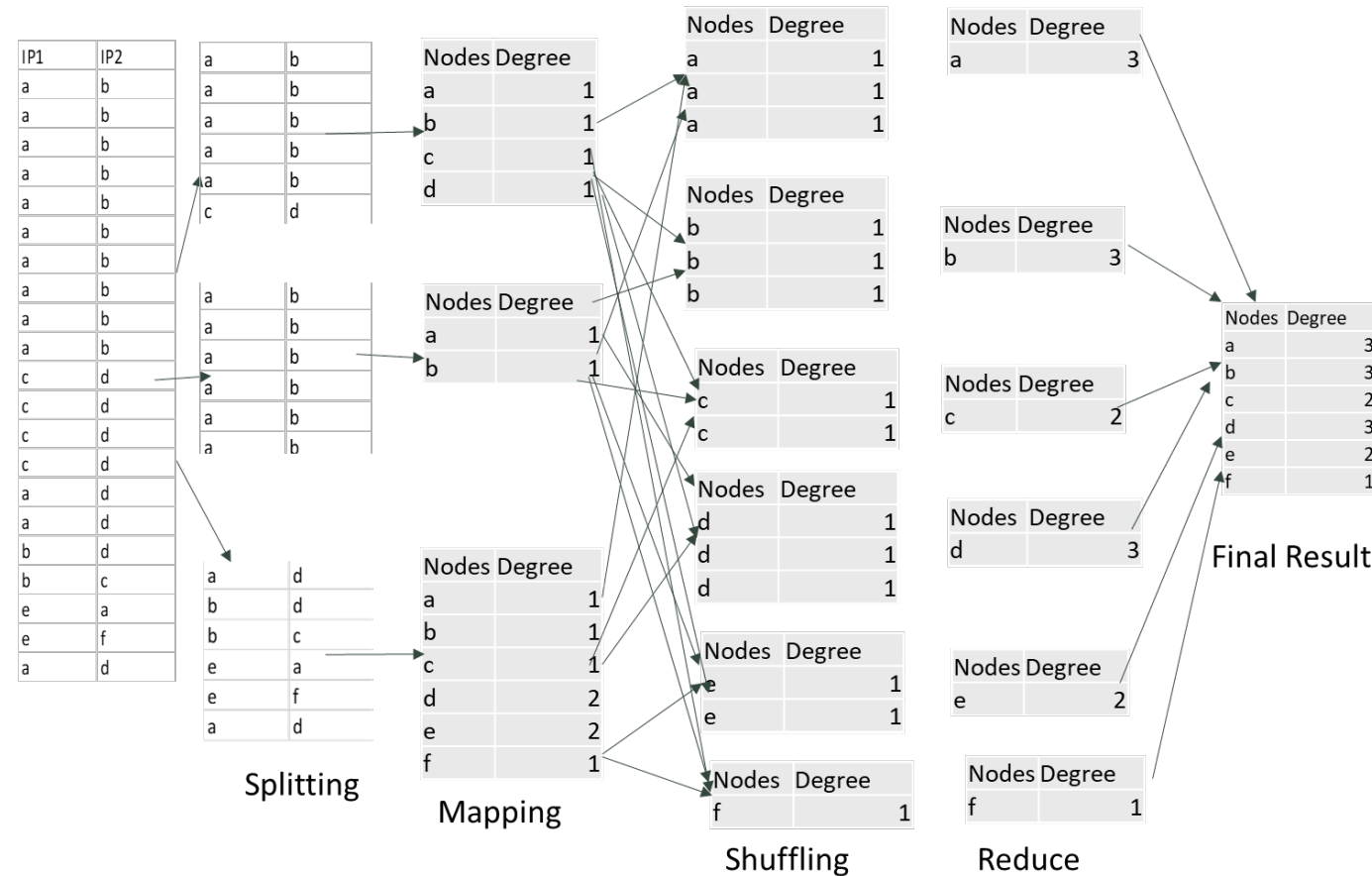
Nodes	Degree T3
a	3
b	1
c	1
d	2
e	2
f	1

Central Nodes a,d,e



# Parallel Graph Analysis

- Let us consider the graph analysis in parallel through big data technologies
- We depict the map reduce process for a single time period
- Here the split can be done within a time period
- Alternatively, the split could be done by time period
- Within a time period we split the data and the degree count is done within those chunks where the key is the node name and the value is the degree, thus forming the key, value pair
- The data is computed for each node in the shuffling and finally the values are tallied for each node in the reduce phase producing the final result
- This mechanism of map reduce is common across several big data tools
- The internal workings of how the data is split and the map reduce is performed may vary



# Multipronged Attacks

---

Multi-pronged attacks are attacks which are spread out over time and several points in the network. Discovering such attacks becomes a challenging task particularly because the datasets become massive and heterogeneous

---

Distributed Intrusion Detection System (dIDS) provides the infrastructure for the detection of a coordinated attack against an organization and its partners' distributed network resources

---

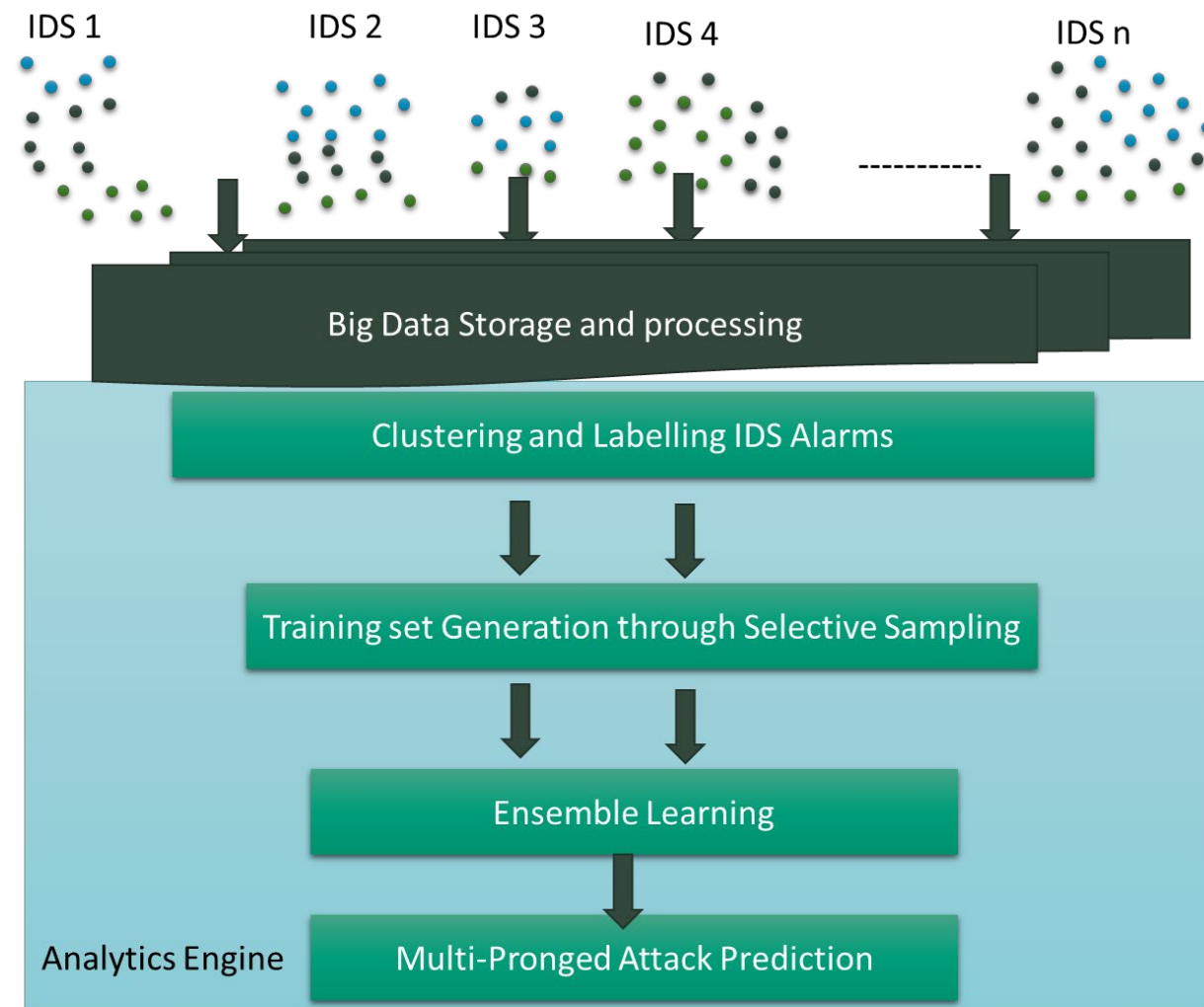
Given the complexity of multiple attack sources and the massive amounts of data generated for such a multi-pronged attack, a multi-level mining framework can be utilized

---

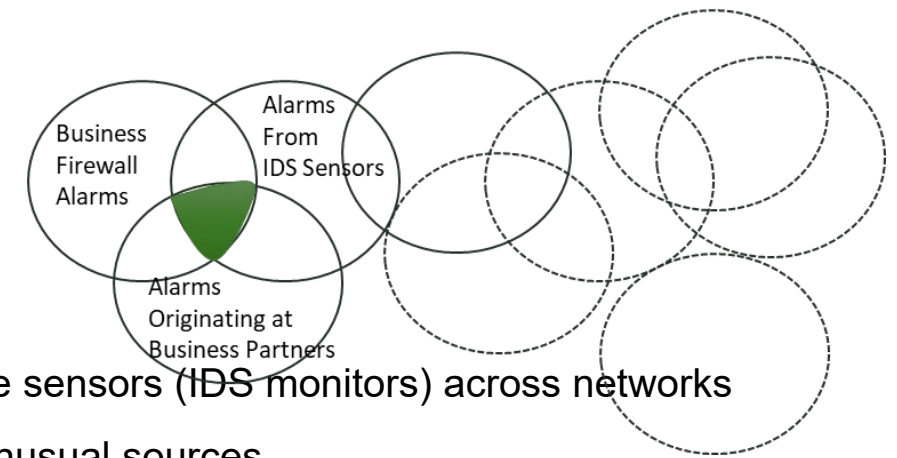
Example architecture utilizes IDS logs to sift through alarms which may look benign individually but may indicate a critical alert in combination with other alerts

# Distributed IDS alarm Scenario

- In this distributed environment within each subnet a dIDS agent performs local intrusion detection and generates IDS log data as IDS1...IDS<sub>n</sub>
- Log data from each dIDS agent is sent to a control center, employing big data tools, where it is stored for aggregated analysis
- Each signature-based agent generates a priority level associated with the alarm when an attack against a known threat is detected, and generates high, medium and low priority alarms for 'anomalous' behavior
- High priority alarms can be clearly labelled, however the low and medium priority alarm data is very large making it difficult to perform manual analysis by an administrator
- In such a scenario, several alarms which are part of a coordinated attack will be missed
- If we can show that the high level alarms have similarities with low level alarms we can propagate the labels to the low level alarms. Once we label them as similar to high level alarms we can try and study them carefully for possible breaches which are part of a coordinated attack.



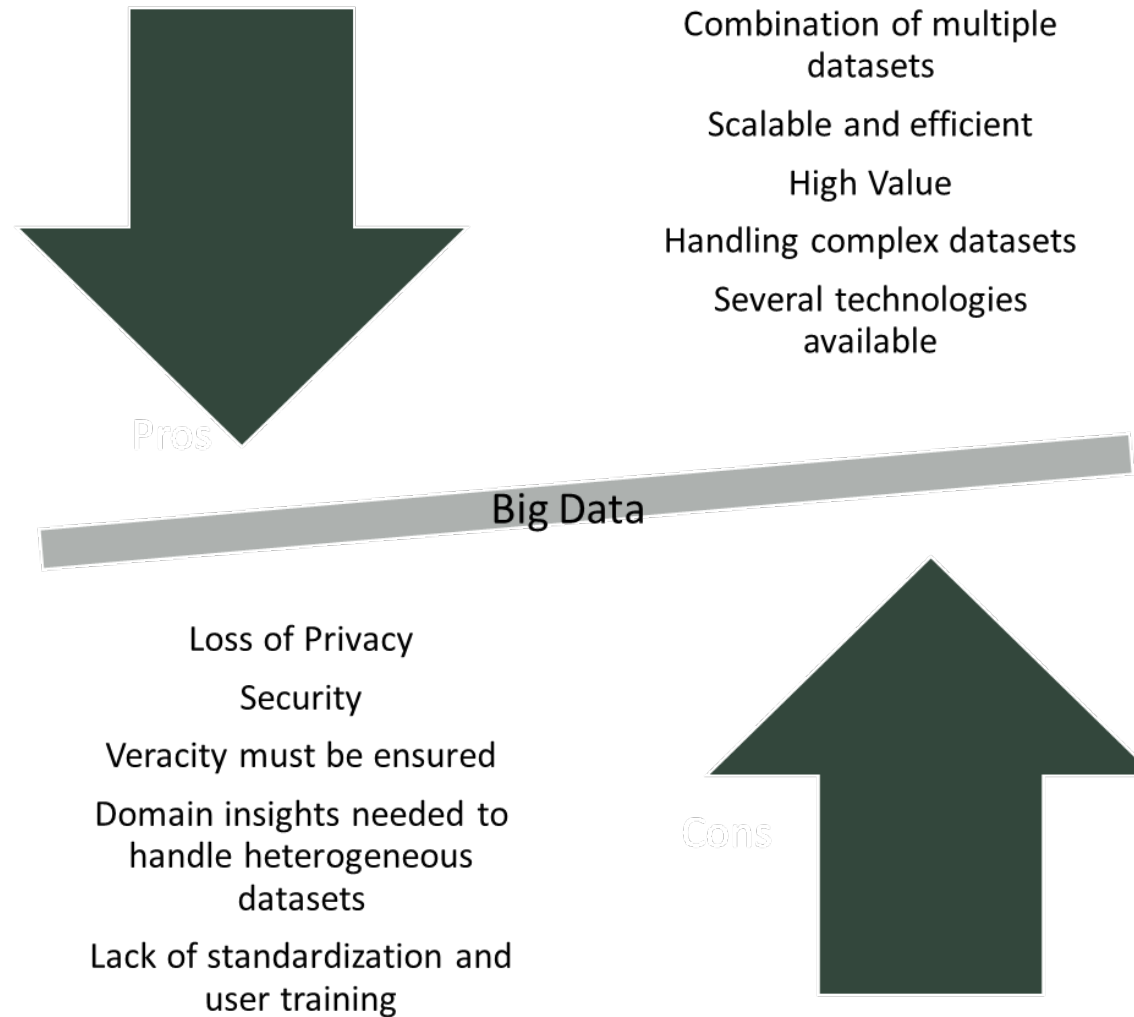
# Alarm Similarity



- There may be several alarms generated at multiple sources or from multiple sensors (IDS monitors) across networks
- These alarms may range from abnormal traffic, unusual port access, and unusual sources
- The key idea here is to connect the anomalies using co-occurrence of alarms at specific time periods, specific location (node/computer) or similar abnormal behavior
- This essentially identifies which have overlaps or similarities in terms of some of the features
- How do we find similarities?
  - This can be done through clustering the alarms together and if alarms fall in the same cluster then the labels from a higher level alarm can be propagated to a lower level alarm
  - Once we have these propagated labels, after domain user validation, we can use this data for future predictions as well
- Let us consider each agent to provide a training set which is generated after preprocessing the data through a clustering algorithm
- Then classification in this data can be seen as a class imbalanced learning problem
- This approach uses an ensemble classification technique to automatically classify the large volume of aggregated alarm data and to alert a system administrator to a potential coordinated attack



# Privacy and Security Issues in Big Data



# References

- Ben Walker, Every Day Big Data Statistics – 2.5 Quintillion Bytes of Data Created Daily, 2015  
<http://www.vcloudnews.com/every-day-big-data-statistics-2-5-quintillion-bytes-of-data-created-daily/>, last accessed 4/12/17
- Ever Merchant, 2014, Ecommerce in Real-Time: How Money is Spent on the Internet  
<http://www.digitalinformationworld.com/2014/07/ecommerce-in-real-time-infographic.html>, last accessed 4/12/17
- (NBDIF), V1.0, 2015, NIST Big Data interoperability Framework,  
[https://bigdatawg.nist.gov/V1\\_output\\_docs.php](https://bigdatawg.nist.gov/V1_output_docs.php), last accessed 4/12/17
- Ward, J. S., & Barker, A. (2013). Undefined by data: a survey of big data definitions. arXiv preprint arXiv:1309.5821.
- Beyer, M. A., & Laney, D. (2012). The importance of 'big data': a definition. Stamford, CT: Gartner, 2014-2018.
- CIA World Factbook, <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html> , last accessed 4/12/17
- Akamai, Q4 2016 State of the Internet / Security Report, 2016  
<https://www.akamai.com/us/en/about/our-thinking/state-of-the-internet-report/global-state-of-the-internet-security-ddos-attack-reports.jsp>, last accessed 4/12/17
- L24, First national cyber security exercise Cyber Shield, 2016,  
<http://l24.it/en/society/item/150489-first-national-cyber-security-exercise-cyber-shield-2016-will-be-held>, last accessed 4/12/17
- Cloud Security Alliance, [Big Data Taxonomy - Cloud Security Alliance, September 2014](https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf), [https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big\\_Data\\_Taxonomy.pdf](https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf), Last accessed 4/13/17.
- [Javiar Roman](https://hadoopecosystemtable.github.io/), The Hadoop Ecosystem Table, <https://hadoopecosystemtable.github.io/> , Last accessed 4/13/17.
- Akshay Grover, [Jay Gholap](#), [Vandana P. Janeja](#), [Yelena Yesha](#), [Raghu Chintalapati](#), [Harsh Marwaha](#), [Kunal Modi](#): SQL-like big data environments: Case study in clinical trial analytics. *Big Data 2015*: 2680-2689
- Tan, Y., Vuran, M. C., & Goddard, S. (2009, June). Spatio-temporal event model for cyber-physical systems. In Distributed Computing Systems Workshops, 2009. ICDCS Workshops' 09. 29th IEEE International Conference on (pp. 44-50). IEEE.
- Keivan Kianmehr, Negar Koochakzadeh: Learning from socio-economic characteristics of IP geo-locations for cybercrime prediction. *IJBIDM* 7(1/2): 21-39 (2012)
- Ferebee, D., Dasgupta, D., & Wu, Q. (2012, December). A cyber-security storm map. In Cyber Security (CyberSecurity), 2012 International Conference on (pp. 93-102). IEEE.
- Banerjee, A., Venkatasubramanian, K. K., Mukherjee, T., & Gupta, S. K. S. (2012). Ensuring safety, security, and sustainability of mission-critical cyber-physical systems. *Proceedings of the IEEE*, 100(1), 283-299.
- K. Venkatasubramanian, S. Nabar, S. K. S. Gupta, and R. Poovendran, Cyber Physical Security Solutions for Pervasive Health Monitoring Systems, M. Watfa, Ed. IGI Global, 2011, ser. E-Healthcare Systems and Wireless Communications: Current and Future Challenges.
- Namayanja, J. M., & Janeja, V. P. (2014, October). Change detection in temporally evolving computer networks: A big data framework. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 54-61). IEEE.
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005, August). Graphs over time: densification laws, shrinking diameters and possible explanations. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 177-187). ACM.
- Janeja, V. P., Azari, A., Namayanja, J. M., & Heilig, B. (2014, October). B-dids: Mining anomalies in a Big-distributed Intrusion Detection System. In Big Data (Big Data), 2014 IEEE International Conference on (pp. 32-34). IEEE.
- Norway, the country where you can see everyone's tax returns, 2016 <https://www.theguardian.com/money/blog/2016/apr/11/when-it-comes-to-tax-transparency-norway-leads-the-field>
- Gopalani, S., & Arora, R. (2015). Comparing apache spark and map reduce with performance analysis using k-means. *International Journal of Computer Applications*, 113(1).
- Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Xin, D. (2016). Mllib: Machine learning in apache spark. *Journal of Machine Learning Research*, 17(34), 1-7.
- Wang, H., Wu, B., Yang, S., Wang, B., & Liu, Y. (2014). Research of decision tree on YARN using MapReduce and Spark. In *World Congress in Computer Science, Computer Engineering, and Applied Computing*.
- Verma, J. P., & Patel, A. Comparison of MapReduce and Spark Programming Frameworks for Big Data Analytics on HDFS. 2016, IJCSC, Volume 7 • Number 2 March 2016 - Sept 2016 pp. 80-84
- Misal, V., Janeja, V. P., Pallaprolu, S. C., Yesha, Y., & Chintalapati, R. (2016, December). Iterative unified clustering in big data. In *Big Data (Big Data), 2016 IEEE International Conference on* (pp. 3412-3421). IEEE.