I. Introduction

Big data and big data analytic elements trace back to the 1600s where people were trying to gather information from large data sets that they had to spend a lot of time on. In the 1600s John Graunt was analyzing the deaths of the Bubonic plague. Then in the 1880s data the US was collecting so much data that it would take them 10 years to process, but the tabulating machine changed that. It was finally in 2001 that Doug Laney coined the term big data and described what it was. From software like Hadoop and Nutch emerged which helps in the processing of big data. Currently, people are using big data for analytics. Big data is only going to continue to grow because of IoT.

II. Early Principles

A. Big Data

In the 1500s London began issuing a weekly report of who died, how they died, and how many people had been born. Gathering and publishing this information began because of recurrences of the bubonic plague. In the 1620s, John Graunt compiled all the information and turned it in tables using analytics. Graunt knew the data was open and lacking analysis. He published a book on his research and findings. He was the first to create life tables for a large population from a significant amount of information. [6] This idea of gathering a vast amount of information and transforming it into other products/recognizing patterns not first apparent is one of the first occurrences of big data and big data analytics. 1880s immigrants were flooding into the United States. Herman Hollerith was a clerk at the US Census Bureau. Hollering had a graduate degree in engineering and when it was suggested that a machine should be able to handle all the data. Hollerith took guidance from Joseph Marie Jacquard who used holes in cardboard cards to guide the pattern for a sewing machine and Charles Babbage who tried building a steam-powered information processor to do polynomial calculation using gears and crankshafts. The machine Hollerith created would require an operator to place a 3 by 7 in card with holes indicating the information of a certain citizen and then a set of pins would go through the holes and complete a circuit if the pins touched mercury then storing the memory. This machine decreased the time needed to process the data for the bureau from 10 years to 2 years. [2] The tabulating machine organizations ability to process information and soon the machine was being used in all sorts of industry like ”department stores, electric and gas utilities, chemical and drug manufacturers, steelworks, oil companies, and railroad”. The tabulating machine demonstrated the data storage characteristic and for decades how efficient data processing of large sets of data could elevate a business.

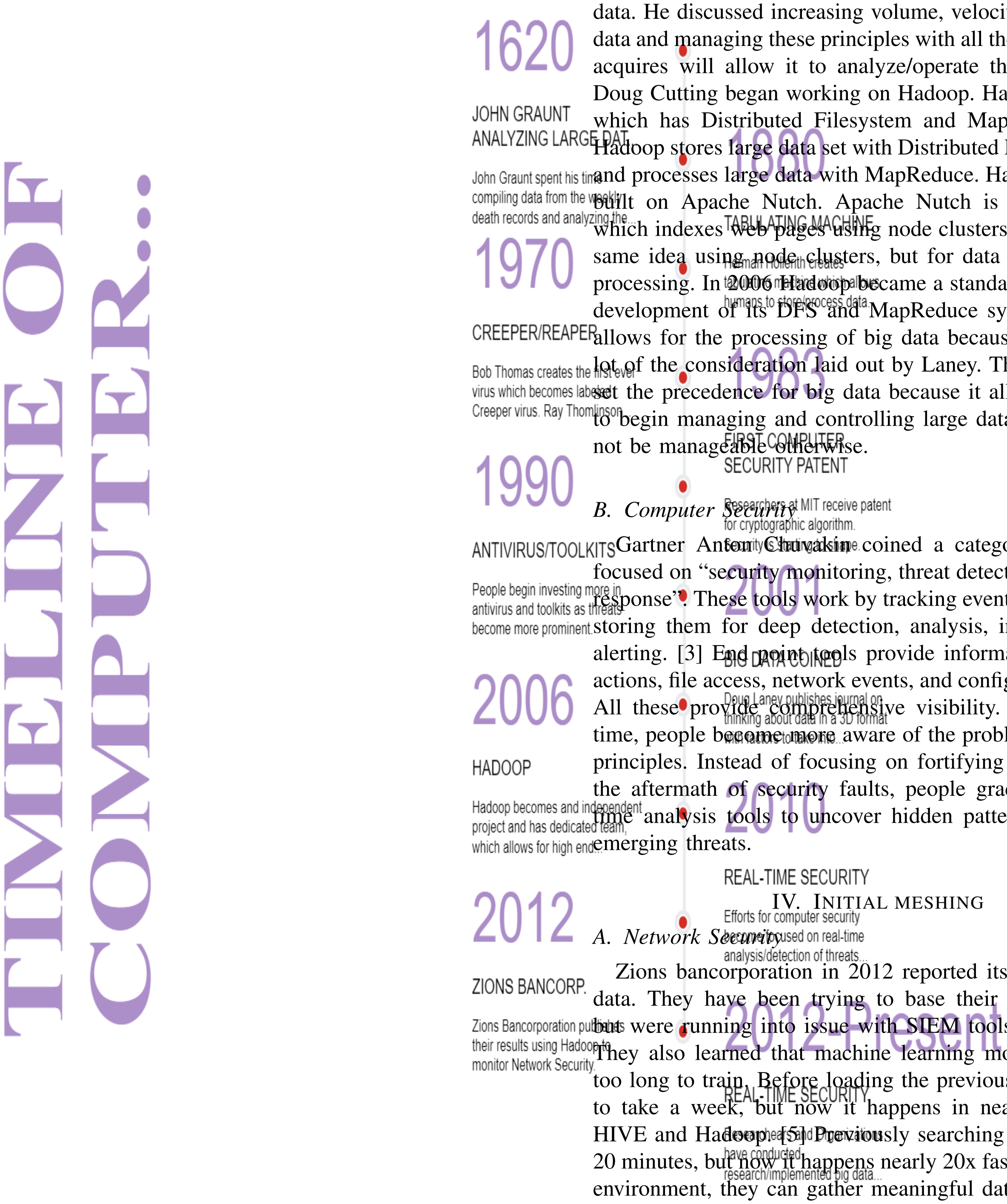
B. Computer Security

An important feature for viruses is that they self-replicate, which was outlined by John con Neumann in 1949. A computer virus that replicates over a network is called a worm. The Creeper worm was created Robert H. Thomas in 1970. The Creeper was written in PDP-10 assembly and ran on Tenex OS, and its only purpose was to demonstrate that a self-replicating automaton was possible therefore it would print “I'M THE CREEPER: CATCH ME IF YOU CAN”. The Creeper virus never actually replicated itself, it would start creating a file on the computer it was on to demonstrate that it could self-replicate, then it would open a connection with another Tenex machine and move itself. With reports of the creeper, Robert H. Thomas' colleague, Ray Tomlinson created the Reaper program, which was similar in structure to the Creeper, but would move through the net and try to remove the occurrences of Creeper on any computer. Although not was widespread, this was the initial proof of concept for cyber threats/security. [1] The creeper virus demonstrated that networks could be exploited, and the Reaper program demonstrated security principles by cleaning up the aftermath of the virus. In 1977 scientists Ron Rivest, Adi Shamir, and Leaonard Adleman at MIT released a paper describing RSA Algorithm. In 1983 they received the first patent for a computer technology having to do with security. Their algorithm generates keys, encrypts the information, and then the information is decrypted. The algorithm uses a public key to encrypt, with only the public key being known. The private key is then used to decrypt the message. [8] At this point in computer security, people are starting to focus on defending their information/data by trying to make it impossible to access unless someone should have access to it. Initially viruses came at a slow rate and so the first antivirus programs to come out only dealt with a small number of viruses. Then in the early 90s “toolkits” were released which would search a computer for an occurrence of a virus. These toolkits could also remove the virus or create a checksum of the clean file to prevent future infections. In the later 90's anti-virus developers created real-time scanners that would monitor the system and check for virus presence. Developers began applying heuristic analysis to look for viruses by identifying similar characteristics to known viruses. This approach with handling viruses/cyber threats would carry over for a long time where the common thing to do was to fortify against any sort of penetration/intrusion and clean up the damage.

III. Evolution/Transfer into Big Data and Security through Big Data

A. Big Data

In 2001 Doug Laney published a paper outlining the struggles organizations were having managing data. Laney explained with the direction that the internet is headed that companies were going to start struggling to manage the data that they were receiving. He outlines condensing data increases “operational, analytical, and collaborative consistencies”, but “changing economic conditions have made this job more difficult. [7] It is then he outlined that organizations were going to have to deal with volume, the depth of data available, by using tiered storage systems, limiting/monitoring data collection, and collecting unique data. Along with volume, he noticed the data velocity, the speed at which data is being collected, is increasing. He highlights minimizing latency by making data


TIMELINE OF COMPUTER...
1620
JOHN GRAUNT
ANALYZING LARGE DAT...
John Graunt spent his time
compiling data from the weekly
death records and analyzing the...
1880
TABULATING MACHINE
Herman Hollerith creates
tabulating machine which allows
humans to store/process data.
1970
CREEPER/REAPER
Bob Thomas creates the first ever
virus which becomes labeled
Creeper virus. Ray Thomlinson
1983
FIRST COMPUTER
SECURITY PATENT
Researchers at MIT receive patent
for cryptographic algorithm.
Security is starting to take...
1990
ANTIVIRUS/TOOLKITS
People begin investing more in
antivirus and toolkits as threats
become more prominent.
2001
BIG DATA COINED
Doug Laney publishes journal on
thinking about data in a 3D format
with factors to take into...
2006
HADOOP
Hadoop becomes and independent
project and has dedicated team,
which allows for high end...
2010
REAL-TIME SECURITY
Efforts for computer security
become focused on real-time
analysis/detection of threats...
2012
ZIONS BANCORP.
Zions Bancorporation publishes
their results using Hadoop to
monitor Network Security.
2012-Present
REAL-TIME SECURITY
Researchers and Organizations
have conducted
research/implemented big data...

accessible where it needs to be and making the transfer of data as direct as it can be and periodic extraction, integration, and reorganization for data management. He claims variety, the inconsistent data semantics will make data management difficult. Therefore, profiling data to resolve inconsistencies, making adapters for acquiring and delivering data, using universal translators, and indexing techniques to relate data will limit inconsistencies in data types. In this paper Doug Laney laid out the initial principles that would later become known as big data. He discussed increasing volume, velocity, and variety of data and managing these principles with all the data a company acquires will allow it to analyze/operate the data. In 2004, Doug Cutting began working on Hadoop. Hadoop is a system which has Distributed Filesystem and MapReduce. Apache Hadoop stores large data set with Distributed FileSystem(DFS) and processes large data with MapReduce. Hadoop is a project built on Apache Nutch. Apache Nutch is a search engine which indexes web pages using node clusters. Hadoop has the same idea using node clusters, but for data storage and data processing. In 2006 Hadoop became a standalone project with development of its DFS and MapReduce system. [4]Hadoop allows for the processing of big data because it deals with a lot of the consideration laid out by Laney. Therefore, Hadoop set the precedence for big data because it allowed for people to begin managing and controlling large data sets that would not be manageable otherwise.

B. Computer Security

Gartner Anton Chuvakin coined a category of tools that focused on "security monitoring, threat detection, and incident response". These tools work by tracking events on the host and storing them for deep detection, analysis, investigation, and alerting. [3] End point tools provide information for process actions, file access, network events, and configuration changes. All these provide comprehensive visibility. At this point in time, people become more aware of the problem with security principles. Instead of focusing on fortifying and cleaning up the aftermath of security faults, people graduated into real-time analysis tools to uncover hidden patterns indicative of emerging threats.

IV. Initial meshing

A. Network Security

Zions bancorporation in 2012 reported its results with big data. They have been trying to base their security on data but were running into issue with SIEM tools being too slow. They also learned that machine learning models would take too long to train. Before loading the previous days, logs used to take a week, but now it happens in near real-time with HIVE and Hadoop. [5] Previously searching could take up to 20 minutes, but now it happens nearly 20x faster. With the new environment, they can gather meaningful data from firewalls, security devices, website traffic, business processes, and day-to-day transactions to ensure their networks are secure. This was one of the first examples where a corporation took in a variety of voluminous data to identify threats in real time using big data architecture.

B. Enterprise Events Analytics

Enterprises are gathering data at rates they cannot handle, therefore HP Labs focused on designing algorithms to provide information that was more actionable. They introduced a large-scale graph interference approach to identify malware-infected hosts and malicious domains. A belief propagation was used to determine likelihood of a malicious host/domain. By testing on HTTP request data set from a large enterprise, DNS requests data set from an ISP, and a intrusion detection system alert data set from enterprises world-wide the researchers were able to establish a high probability of detection.

C. NetFlow Monitoring to ID Botnets

The Hadoop's MapReduce was used to identify infected hosts participating in a botnet because there are large amounts of NetFlow data. Researchers used BotTrack to examine relationships between hosts to track command-and-control channels. The detection works by creating a dependency graph, running the PageRank algorithm, and DBScan clustering. The PageRank algorithm is the most intensive part so it was implemented using a cluster of 12 commodity nodes. They analyzed a data set containing 16 million hosts and 720 million records. The cluster reduced time by a factor of seven since scores are propagated through edges.

D. APT Detection

APT attack target high-value assets or physical streams. APTs are carried out by highly skilled and well-funded attackers. Because of how specialized these attacks are, they require a lot of individual assets and is not scalable. RSA Labs has observed that the attacker's action causes the user's actions to deviate from their normal pattern. At RSA they are using behavioral deviations to act as anomaly sensors where they can track abnormal behavior leading to an investigation or reports of a given user's activity. Beehive can process 1 billion logs in about an hour and identify policy violations/malware infections that would not have been noticed.

V. Conclusion

References

[1] First computer virus.
[2] The punched card tabulator.
[3] Lital Asher-Dotan. Endpoint detection and response (edr) 101.
[4] Marko Bonaci. The history of hadoop, Nov 2018.
[5] Alvaro A Cardenas. Big data for security intelligence. Sep 2013.
[6] Kat Eschner. People have been using big data since the 1600s, Apr 2017.
[7] Doug Laney. Application delivery strategies. Feb 2001.
[8] Paul Tarau. Rsa(cryptosystem). Oct 2014.