BiG Data Security

CISC 6640 PRIVACY AND SECURITY IN BIG DATA

Instructor:
Md Zakirul Alam Bhuiyan
Assistant Professor
Department of Computer and Information Sciences
Fordham University

We Have Learned …

- Database Security
- Relational Databases
  - Database security models
- No SQL Databases
- Object Based vs. Object Oriented
- Overview of Database Vulnerabilities
  - Common DBMS vulnerabilities
- Overview of Database topics/issues (indexing, inference, aggregation, polyinstantiation)
  - Security issues of inference and aggregation
- Hashing and Encryption
- Database access controls (DAC, MAC, RBAC, Clark-Wilson)
- Information flow between databases/servers & applications

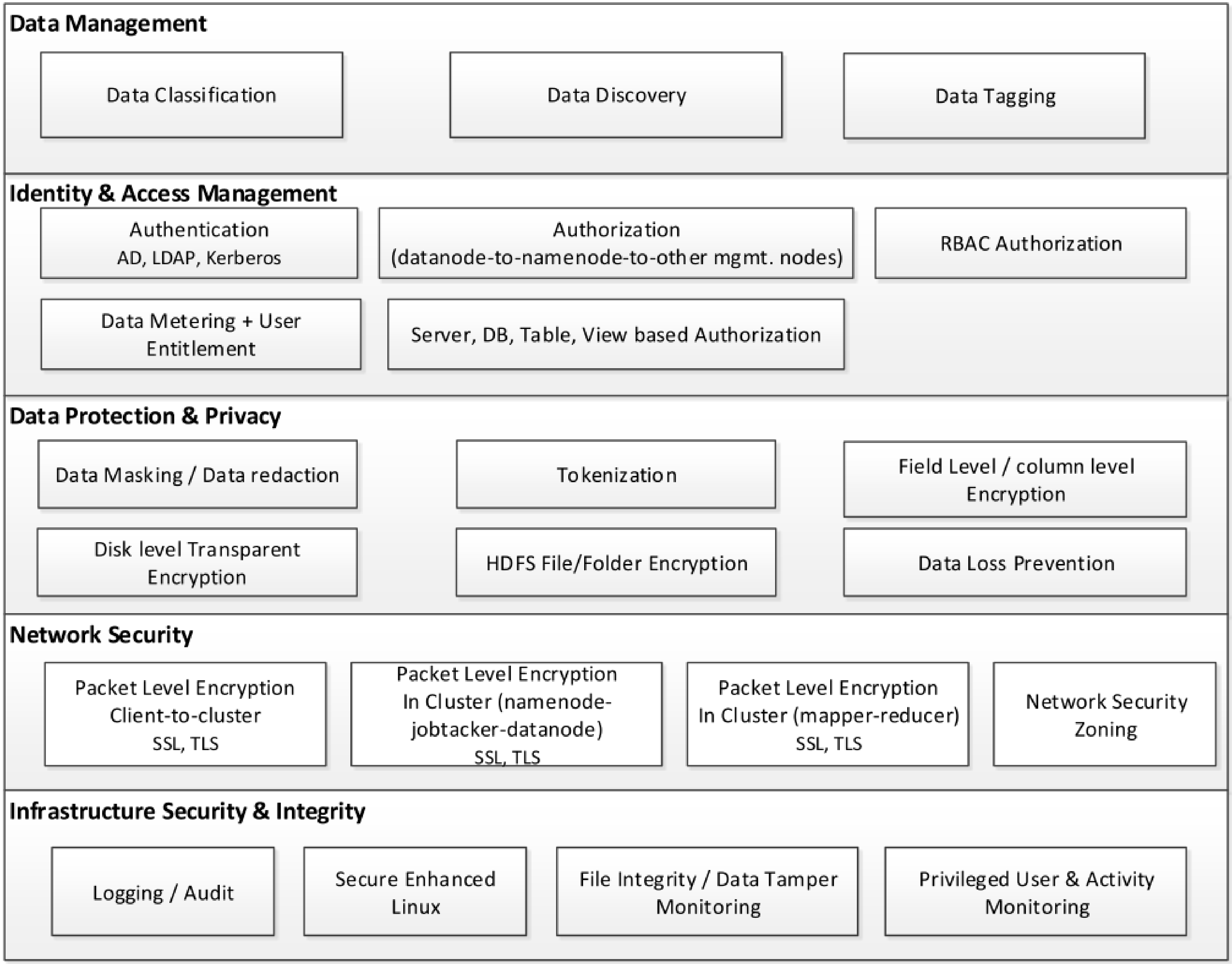
What We Are Going to Learn…

- Big Data Security Framework
  - Data Management
  - Identity & Access Management
  - Data Protection & Privacy
  - Network Security
  - Infrastructure Security & Integrity

Big Data Security Framework

- The '5 pillars' of big data security framework:
  1. Data Management
  2. Identity & Access Management
  3. Data Protection & Privacy
  4. Network Security
  5. Infrastructure Security & Integrity

They are further decomposed into 21 sub-components, each of which are critical to ensuring the security and mitigating the security risk and threat vectors to the Big Data stack



Data Management

- Data Classification
- Data Discovery
- Data Tagging

Identity & Access Management

- Authentication
AD, LDAP, Kerberos
- Authorization
(datanode-to-namenode-to-other mgmt. nodes)
- RBAC Authorization
- Data Metering + User Entitlement
- Server, DB, Table, View based Authorization

Data Protection & Privacy

- Data Masking / Data redaction
- Tokenization
- Field Level / column level Encryption
- Disk level Transparent Encryption
- HDFS File/Folder Encryption
- Data Loss Prevention

Network Security

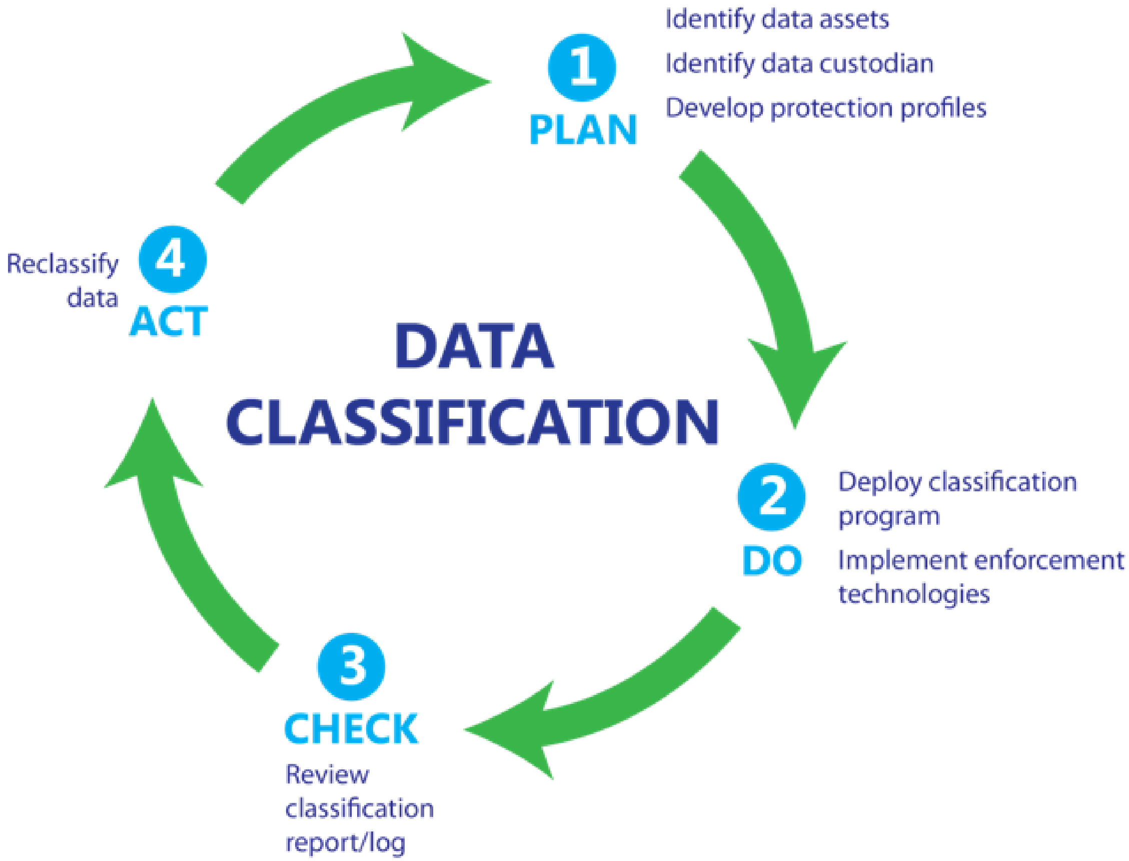
- Packet Level Encryption
Client-to-cluster
SSL, TLS
- Packet Level Encryption
In Cluster (namenode-jobtacker-datanode)
SSL, TLS
- Packet Level Encryption
In Cluster (mapper-reducer)
SSL, TLS
- Network Security Zoning

Infrastructure Security & Integrity

- Logging / Audit
- Secure Enhanced Linux
- File Integrity / Data Tamper Monitoring
- Privileged User & Activity Monitoring

Data Management

- Data Classification



DATA CLASSIFICATION

1 PLAN

Identify data assets

Identify data custodian

Develop protection profiles

2 DO

Deploy classification program

Implement enforcement technologies

3 CHECK

Review classification report/log

4 ACT

Reclassify data

Data Management

- Data Classification
  - **Determine all distinct data fields**
    - Work with your legal, privacy office, intellectual property, finance, and information security.
  - **Perform a security control assessment exercise**
    - Determine location of data
      - e.g. exposed to internet, secure data zone
    - Determine number of users and systems with access
    - Determine security controls
      - e.g. can it be protected cryptographically

Data Management

- Data Classification
  - **Determine value of the data to the attacker**
    - Is the data easy to resell on the black market?
    - Do you have valuable Intellectual Property (e.g. a nation state looking for nuclear reactor blueprints)
  - **Determine compliance and revenue impact**
    - Determine breach reporting requirements for all the distinct fields
    - Does loss of a particular data field prevent you from doing business
      - e.g. card holder data
    - Estimate re-architecting cost for current systems
      - e.g. buying new security products

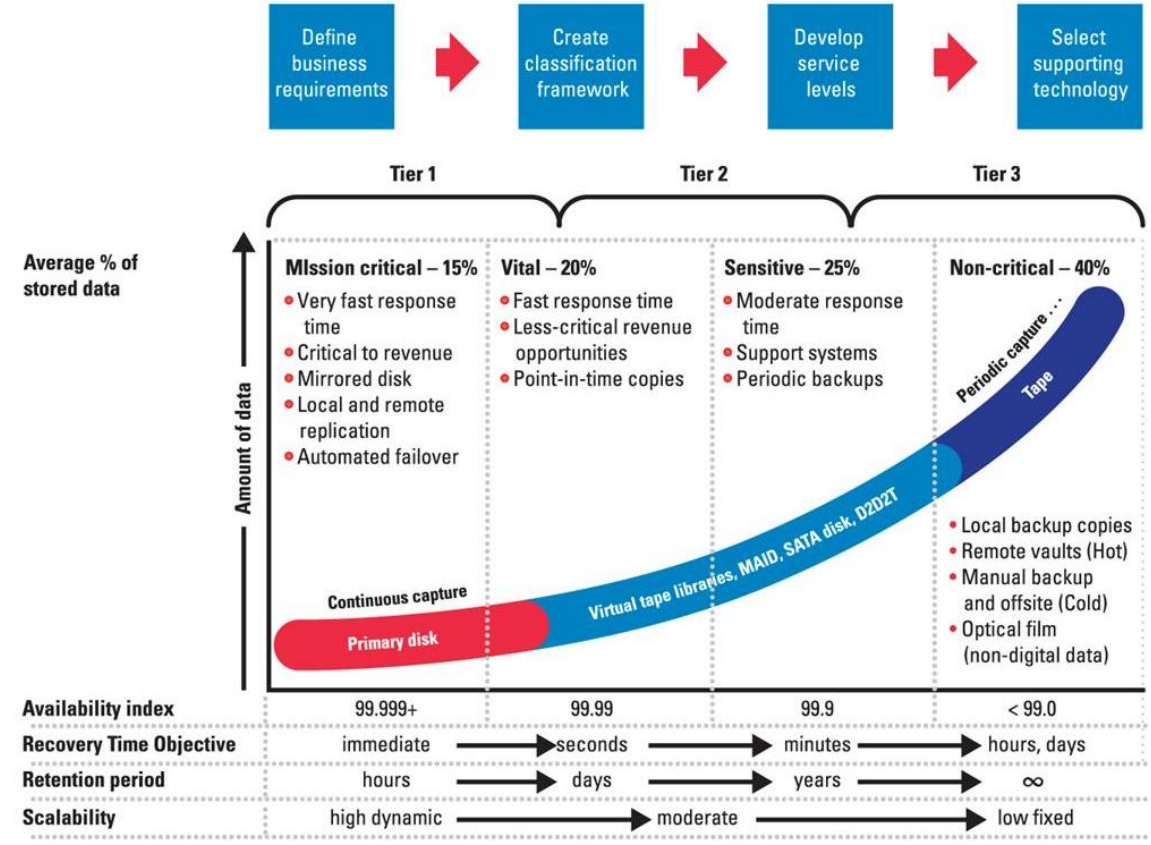
Data Management

- Data Classification
  - **Determine impact to the owner of the PII data, e.g. a customer**
    - Does the field cause phishing attacks, e.g. email vs. just replace it e.g. loss of a credit card

Data Management

Data Classification

- Data classification model



Define business requirements → Create classification framework → Develop service levels → Select supporting technology

Tier 1 / Tier 2 / Tier 3

Average % of stored data

MIssion critical – 15%
- Very fast response time
- Critical to revenue
- Mirrored disk
- Local and remote replication
- Automated failover

Vital – 20%
- Fast response time
- Less-critical revenue opportunities
- Point-in-time copies

Sensitive – 25%
- Moderate response time
- Support systems
- Periodic backups

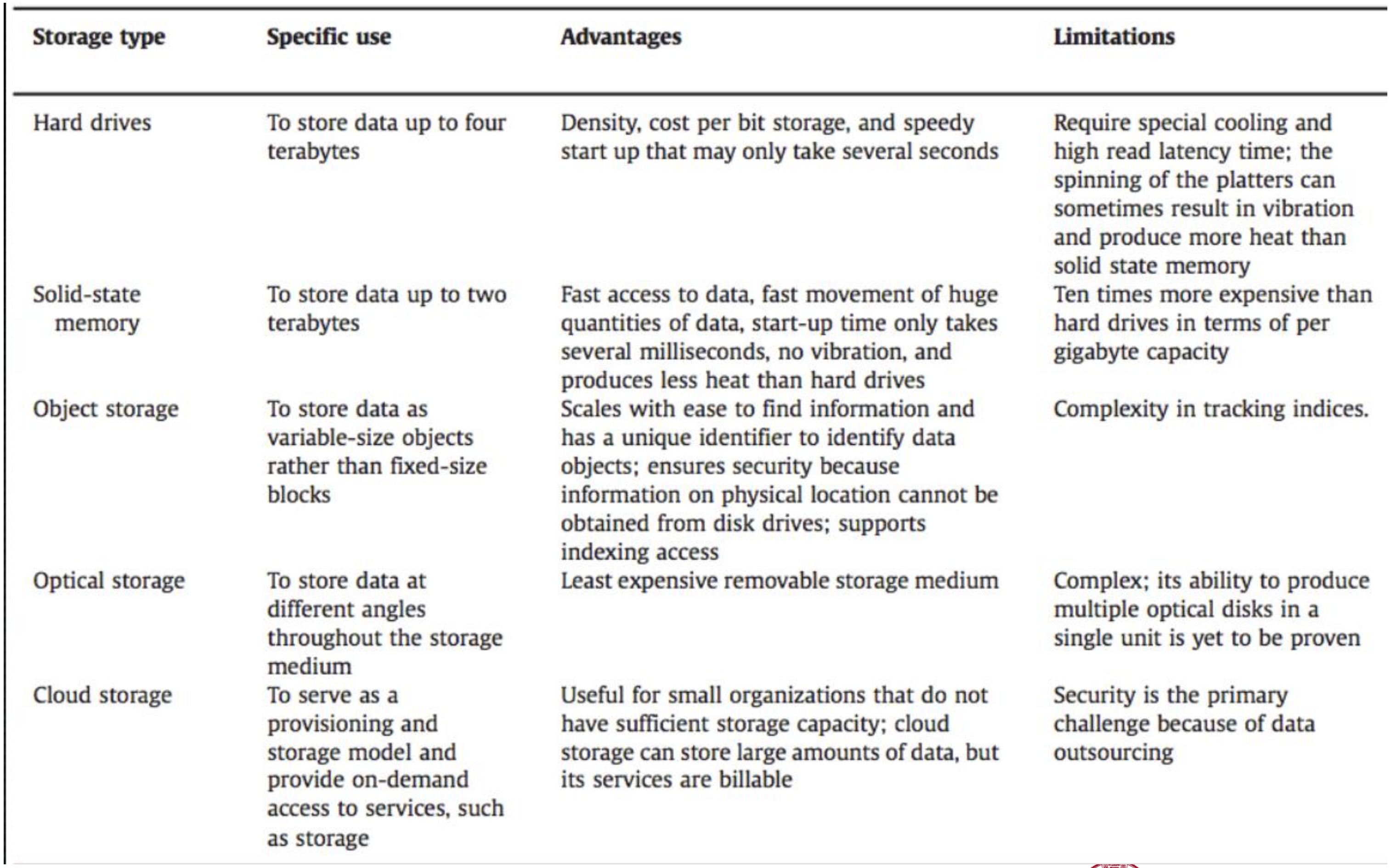
Non-critical – 40%
- Local backup copies
- Remote vaults (Hot)
- Manual backup and offsite (Cold)
- Optical film (non-digital data)

Availability index	99.999+	99.99	99.9	< 99.0
Recovery Time Objective	immediate →	seconds →	minutes →	hours, days
Retention period	hours →	days →	years →	∞
Scalability	high dynamic →	moderate →		low fixed

Source: Horison Information Strategies

Data Management

○ Data Classification

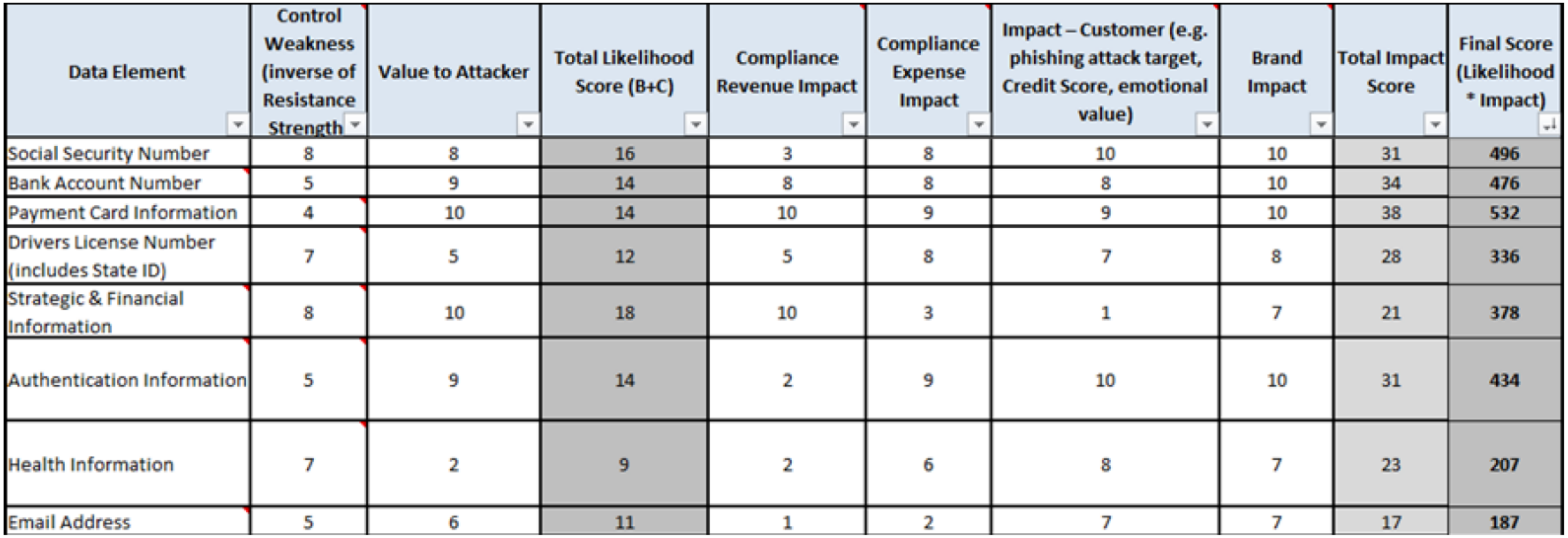


Storage type	Specific use	Advantages	Limitations
Hard drives	To store data up to four terabytes	Density, cost per bit storage, and speedy start up that may only take several seconds	Require special cooling and high read latency time; the spinning of the platters can sometimes result in vibration and produce more heat than solid state memory
Solid-state memory	To store data up to two terabytes	Fast access to data, fast movement of huge quantities of data, start-up time only takes several milliseconds, no vibration, and produces less heat than hard drives	Ten times more expensive than hard drives in terms of per gigabyte capacity
Object storage	To store data as variable-size objects rather than fixed-size blocks	Scales with ease to find information and has a unique identifier to identify data objects; ensures security because information on physical location cannot be obtained from disk drives; supports indexing access	Complexity in tracking indices.
Optical storage	To store data at different angles throughout the storage medium	Least expensive removable storage medium	Complex; its ability to produce multiple optical disks in a single unit is yet to be proven
Cloud storage	To serve as a provisioning and storage model and provide on-demand access to services, such as storage	Useful for small organizations that do not have sufficient storage capacity; cloud storage can store large amounts of data, but its services are billable	Security is the primary challenge because of data outsourcing

Data Management

Data Classification

- **Data Classification Matrix**



Data Element	Control Weakness (inverse of Resistance Strength	Value to Attacker	Total Likelihood Score (B+C)	Compliance Revenue Impact	Compliance Expense Impact	Impact – Customer (e.g. phishing attack target, Credit Score, emotional value)	Brand Impact	Total Impact Score	Final Score (Likelihood * Impact)
Social Security Number	8	8	16	3	8	10	10	31	496
Bank Account Number	5	9	14	8	8	8	10	34	476
Payment Card Information	4	10	14	10	9	9	10	38	532
Drivers License Number (includes State ID)	7	5	12	5	8	7	8	28	336
Strategic & Financial Information	8	10	18	10	3	1	7	21	378
Authentication Information	5	9	14	2	9	10	10	31	434
Health Information	7	2	9	2	6	8	7	23	207
Email Address	5	6	11	1	2	7	7	17	187

Data Management

- Data Discovery
  - **The lack of situational awareness**
    - With respect to sensitive data could leave an organization exposed to significant risks
    - Identifying whether sensitive data is present in Hadoop
      - Where it is located and subsequently triggering the appropriate data protection measures
        - Such as data masking, data redaction, tokenization or encryption is key

Data Management

- Data Discovery
  - **Items are crucial for an effective data discovery exercise of Big Data environment**
    - Define and validate the data structure and schema. This is all useful prep work for data protection activities later
    - Collect metrics (e.g. volume counts, unique counts etc.).
      - For example, if a file has 1M records but it is duplicate of a single person, it is a single record vs. 1M records.
      - This is very useful for compliance but more importantly risk management.

Data Management

- Data Discovery
  - **Items are crucial for an effective data discovery exercise of Big Data environment**
    - Share this insight with your Data Science teams for them to build threat models, profiles which will be useful in data exfiltration prevention scenarios.
    - Build conditional search routines (e.g. only report on date of birth if a person's name is found or Credit Card # + CVV or CC +zip)
    - Account for usecases where once sensitive data has been cryptographically protected
      - e.g. encrypted or tokenized), what is the usecase for the discovery solution.

Coming Attraction…

- Specific Security and Privacy Issues in Big Data