**Gartner Data & Analytics Summit**
Summit 2018
**22 – 23 May 2018 / São Paulo, Brazil**

# Lifeguard Not on Duty: How to Make Your Data Lake Safe and Secure

Claudio Neiva
Merv Adrian

# Who's helping you keep your data lake safe and secure?





Gartner

# Key Issues

1. Why is securing the data lake important?
2. What are the key security elements?
3. How can you make your data lake more secure?

# Key Issues

1. Why is securing the data lake important?
2. What are the key security elements?
3. How can you make your data lake more secure?




Gartner

# Definitions First: What Do You Mean by a Data Lake?

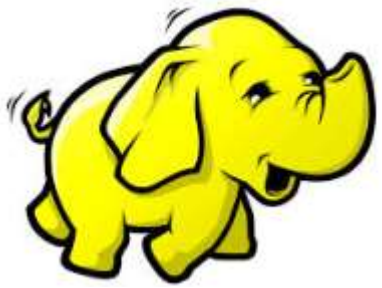
- A collection of storage instances of various data assets additional to the originating data sources.
- These assets are stored in a near-exact or even exact, copy of the source format.
- Presents an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques.

https://www.gartner.com/it-glossary/data-lake



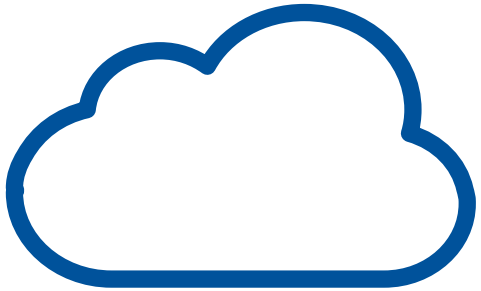

Gartner

# Common Data Lake Implementation Technologies



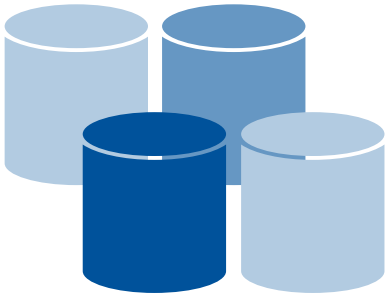
**Hadoop distributions:**

- Simplified data ingestion and storage with several processing options
- Data lake management ecosystem emerging
- Complex deployment and management



**Cloud-based block and object stores:**

- Simplified data ingestion and storage
- Bring your own processing
- Nascent management and security ecosystem



**Database management systems:**

- Optimal for certain data types and formats
- Data processing options expanding beyond SQL
- Scaling and cost may be challenges

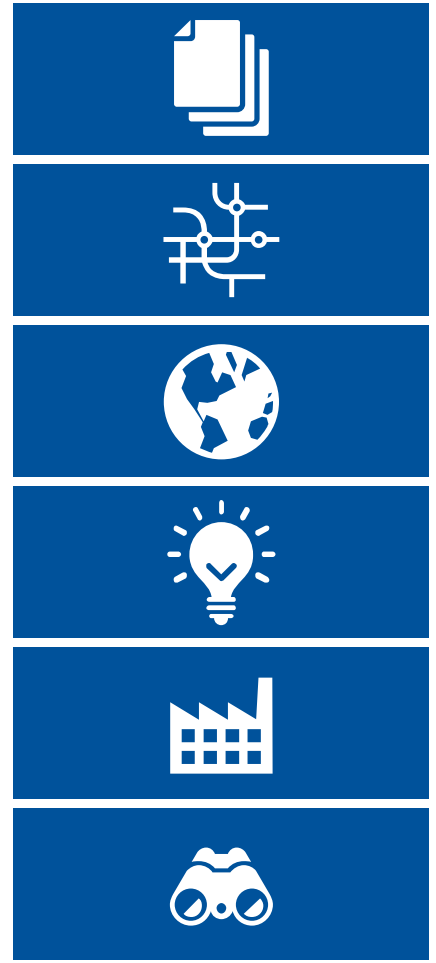



Gartner

# History Is Repeating Itself, but Bigger!

# Security and Risk Concerns About Data Lakes

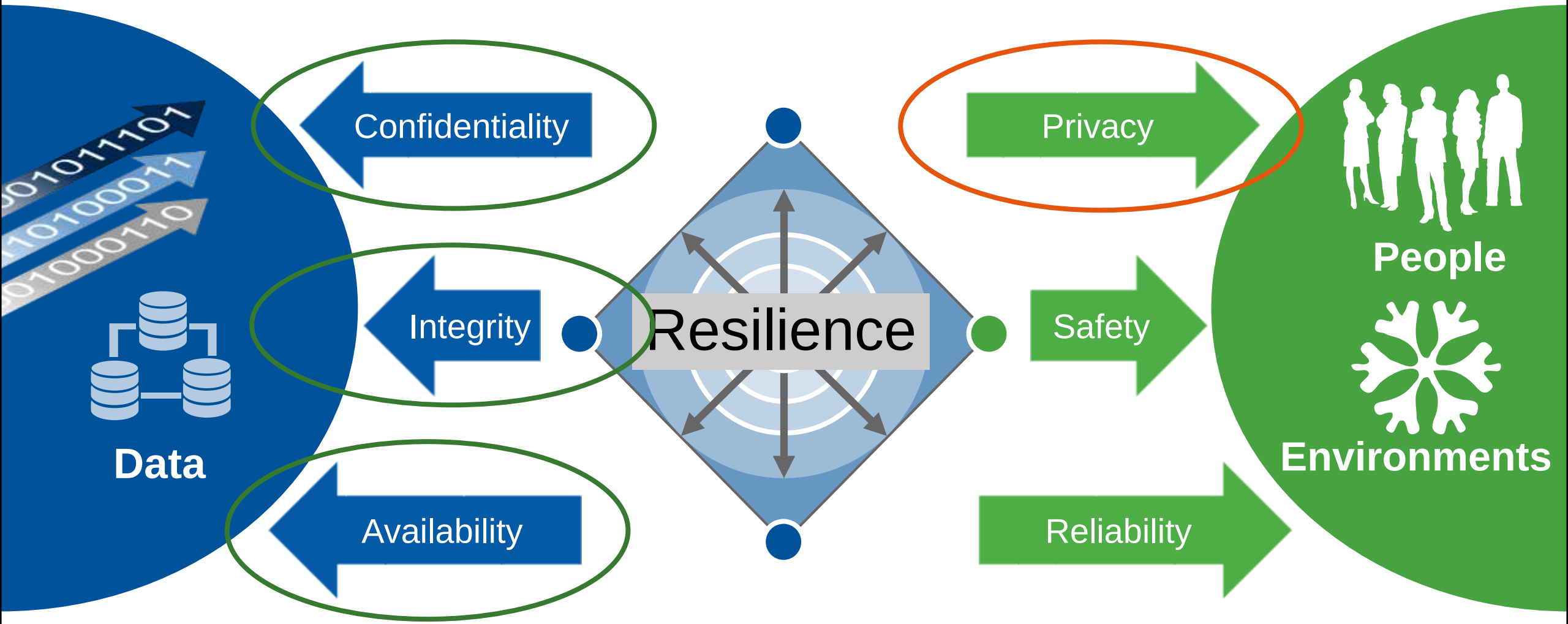
- Data lakes contain LOTS of data.
- Data flows in, through, and out.
- Where is the data lake located?
- Data is transformed and new data is created.
- What infrastructure is supporting the data lake?
- How is the data lake being governed?






Gartner

# Risk and Resilience for the Digital Business



# Strategic Planning Assumptions

- Through 2020, fewer than 30% of Hadoop deployments will be secured and governed in accordance with the enterprise's information governance standards.
- Through 2020, 80% of sensitive data loss incidents from Hadoop will be caused by misadministration, misconfiguration and mismanagement.

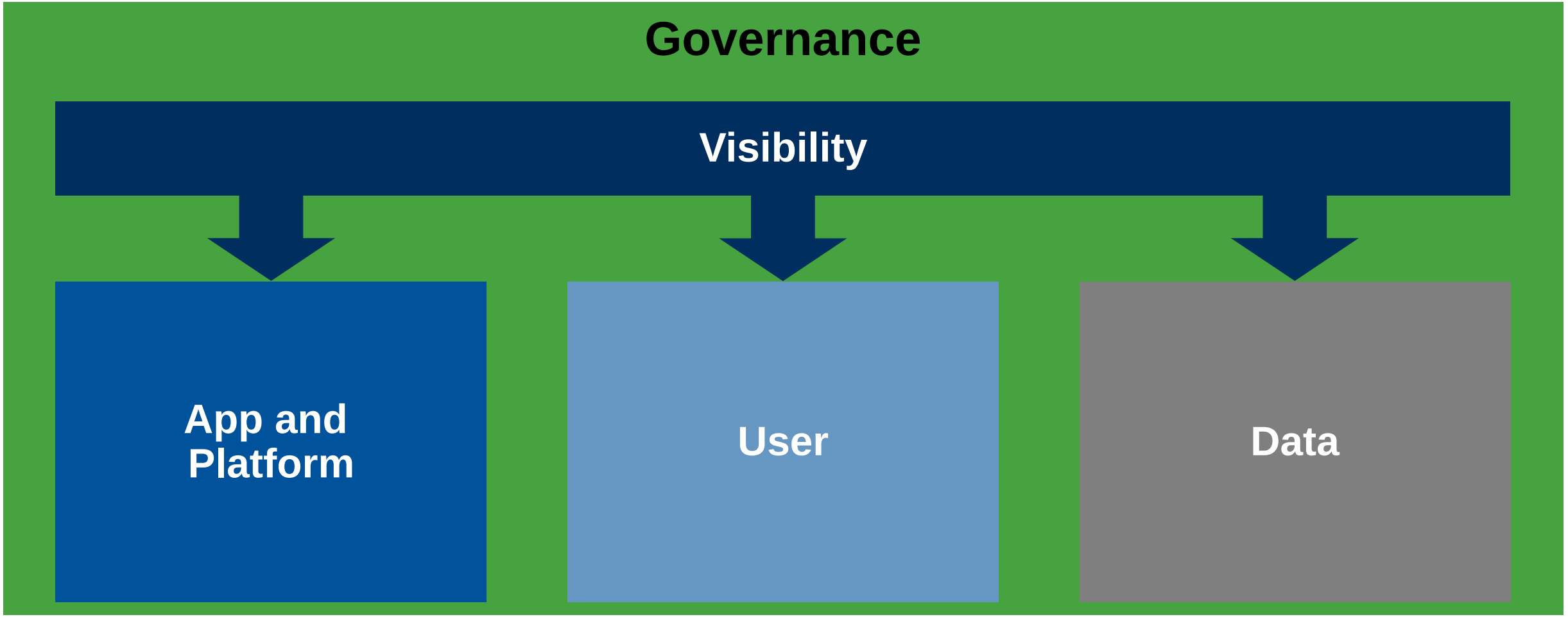
# Key Issues

1. Why is securing the data lake important?
2. What are the key security elements?
3. How can you make your data lake more secure?




Gartner

# Elements of Data Lake Security

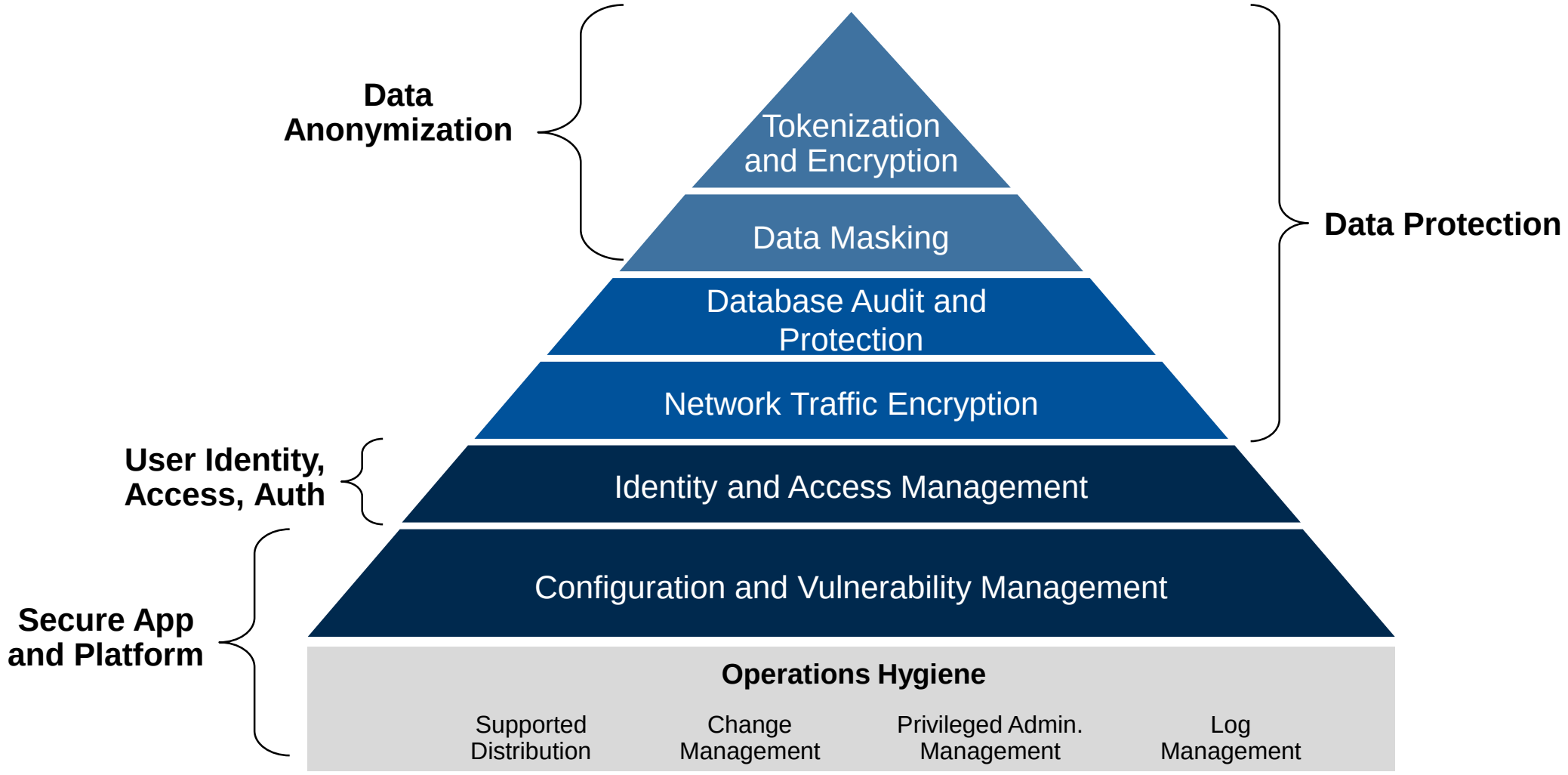






Gartner

# An Architectural View







Gartner

# Security and Risk Concerns of the Apps and Platform

- Inconsistent configurations
- Vulnerabilities not patched
- Lack of product currency
- No centralized privileged administration
- Poor change management
- No audit or event logs generated or centrally collected
- Cloud services are used without sufficient understanding




Gartner

# Managing Users Is a Critical Component

- Controlling the access of a user, what they are allowed to do, and capturing the audit trail of activity is critically important.
- Data lakes may be an island unto themselves, lacking consistent, central management for accounts and access.
- Identity and access management systems are not used.
- Inappropriate levels of authorization or everyone has the same wide-open access.
- No accounting or auditing of user activities.

# Data Security is Not Magic

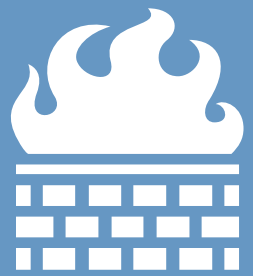



Gartner

# Why Is a Data Security Critical? GDPR Is a Good Reason

The European General Data Protection Regulation (GDPR) will globally impact the processing of all personal data on EU residents and takes effect on 25 May 2018.

**Fines. Yes.**
**Up to 4% of global annual turnover or €20M.**



**Distrusting clients and disloyal employees after breach.**

**Reputation and client loss when disregarding subject's rights.**




Gartner

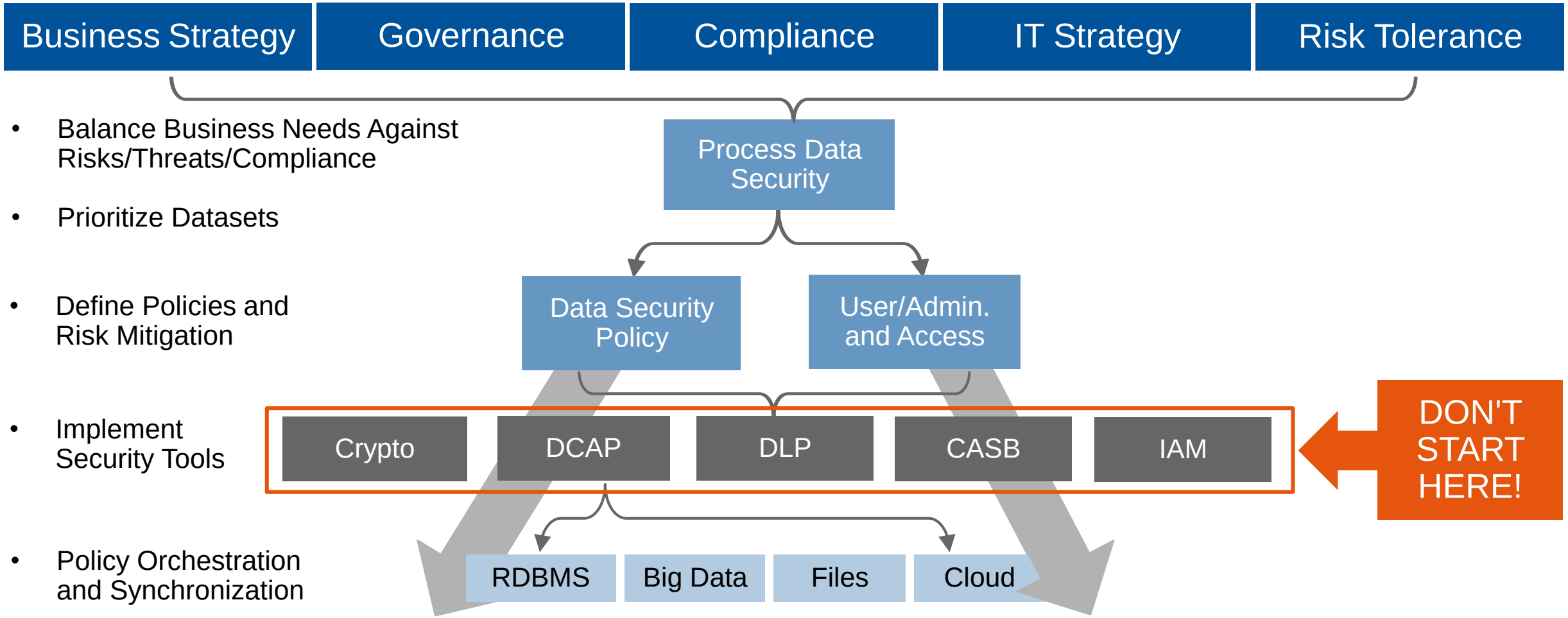
# Visibility

- Are you auditing of all activities being centrally watched?
- Are audit and event logs being provided to the security team? Is there centralized monitoring or is it decentralized/federated?
- Is responsibility agreed for monitoring the different components – platform, application, user, and data?
- Has governance oversight been implemented?




Gartner

# Data Security Governance Gets Organized




- Balance Business Needs Against Risks/Threats/Compliance
- Prioritize Datasets
- Define Policies and Risk Mitigation
- Implement Security Tools
- Policy Orchestration and Synchronization




Gartner

# Key Issues

1. Why is securing the data lake important?
2. What are the key security elements?
3. How can you make your data lake more secure?




Gartner

# It's a Risk Management Issue, So Think About the Risks!

- People are only human and will make mistakes.
- Technologies will fail or not work as expected.
- Increases in system complexity can create larger attack surfaces.
- Motivated attackers are clever and persistent (both insiders and external actors).

Photo Credit: Hiart




Gartner


Security is everyone’s responsibility

# Technologies to Leverage

**User**
Identity Access and Management
Active Directory

**Data Protection**
Encryption
Masking
Tokenization

**Visibility**
SIEM
UEBA

**App and Platform**
Vulnerability Assessment
Apache Ranger
Apache Knox

# Recommendations

- ✓ Have a risk-oriented mindset when building, operating and using a data lake.
- ✓ Leverage your organization's security and risk management expertise, and security tools and capabilities.
- ✓ Use commercially supported solutions and services.
- ✓ All access and activities should be logged and monitored — including all end-user and administrative access.
- ✓ Align to your organization's data governance policies and practices.




Gartner

# Recommended Gartner Research

- **Defining the Data Lake**
  Nick Heudecker, Mark A. Beyer and Lakshmi Randall (G00276838)
- **Securing the Big Data and Advanced Analytics Pipeline**
  Joerg Fritsch (G00313004)
- **Market Guide for Data-Centric Audit and Protection**
  Brian Lowans, Marc-Antoine Meunier and Others (G00298197)
- **Hype Cycle for Data Security, 2017**
  Brian Lowans (G00314204)

For information, please contact your Gartner representative.




Gartner