




The background is a solid dark brown color. In the four corners, there are decorative white line art elements that resemble circuit traces or data paths. These lines connect to small white circles, some of which are arranged in a grid-like pattern. The lines are thin and white, contrasting with the dark background.

MODULE 5A: AUDITING BIG DATA AND DATA REPOSITORIES



OBJECTIVES

- How to audit data repositories
 - Specific audit considerations for big data environments
- 
- 
- 

WHAT IS BIG DATA?


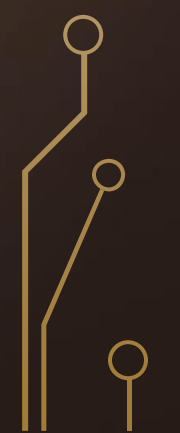
- “Big data” refers to very large and often disparate data sets that are difficult or impossible to process with traditional data handling and analysis solutions. In the case of big data, “very large” can mean data sets up to the exabyte scale (1 exabyte = 1,000 petabytes = 1,000,000 terabytes), although the same technologies can be applied effectively to much smaller data sets.
- Big data may involve a combination of data sources, typically pulled together into a repository designed for consumption of large data sets.

BIG DATA: THE THREE V'S

- Big data is widely described using three key characteristics, or “vectors.” Dubbed the “three Vs” of big data, they were first outlined by Gartner in 2001 and describe the primary ways in which big data differs from traditional data. The three Vs are
 - Volume: describes the large amount of data available or collected
 - Velocity: describes the rapid speed at which data is generated or collected
 - Variety: describes the diversity of data types collected



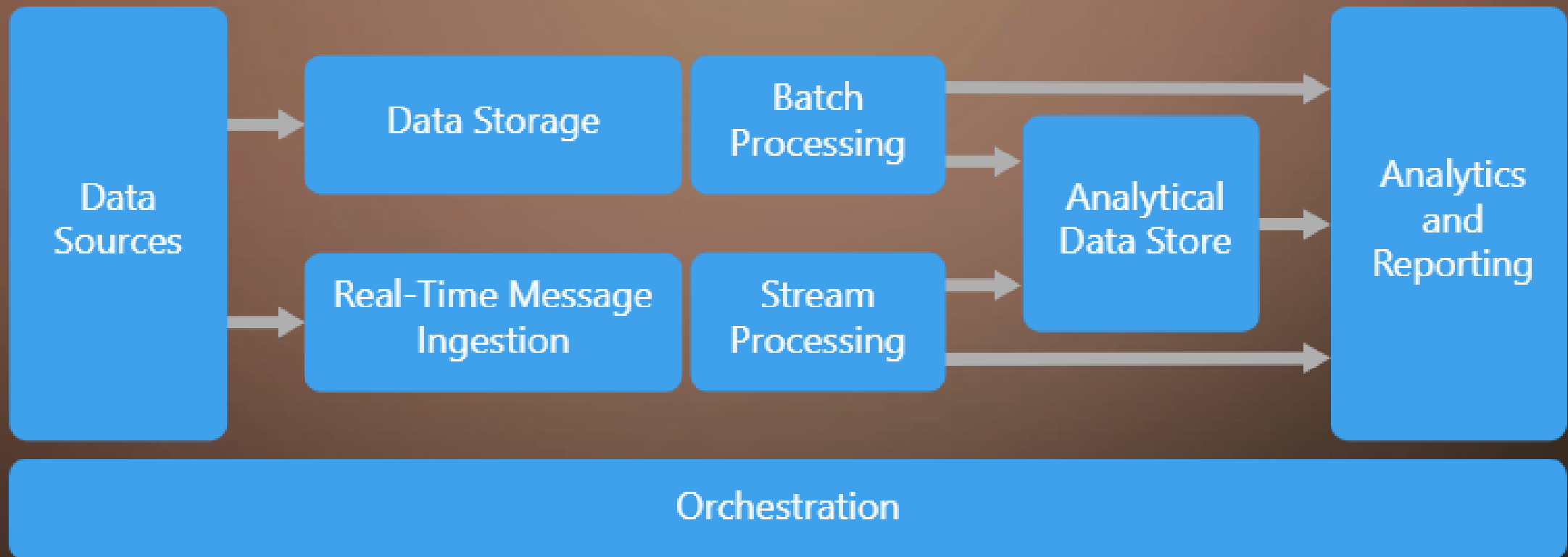
BIG DATA AND DATA REPOSITORY AUDITING ESSENTIALS

- With databases, file servers, big data systems, and a myriad of web-based technologies, there's a lot to consider in the data space. Auditing the entire landscape of data storage and data management in even a medium-sized environment would be a daunting task.
 - Remember to scope your audit carefully with specific goals and target environments in mind.
 - Even with a small scope, the size of a data platform can be intimidating; keep in mind that you can always break the audit down into smaller, more manageable steps.
- 
- 

BIG DATA AND DATA REPOSITORY AUDITING ESSENTIALS

- You will want to familiarize yourself with the various data repositories in use in your business. Leverage your relationships with other IT teams, particularly operations-related groups, to help identify major data systems, both by prevalence and by criticality. Once you've identified the major repositories in play, you can begin to prioritize the list and determine where to focus your audit.
- If you are auditing a data repository that includes file shares, you will need to understand what type of system is serving the files. This could be a Windows or Linux-based server, or a storage system configured to serve file shares. This will help you determine what type of operating system is in use. If the repository is a storage system, it's a safe assumption that the system uses some variant of the Linux operating system.


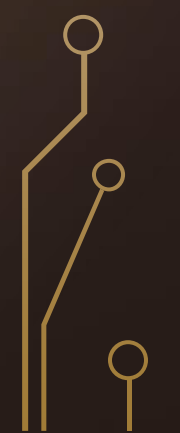
BIG DATA ARCHITECTURE



<https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data>



HOW BIG DATA DIFFER FROM SERVER AUDITING?

- Scale of the data, systems, and storage
 - Distributed across multiple systems, data centers
 - Tie-in to multiple applications, systems
- 
- 

BIGGEST CHALLENGE – DATA SENSITIVITY AND CLASSIFICATION

- **Ensure that the data classification of the environment is understood and review the data ownership process for the environment.**
 - All data stored by or used by a data repository platform should be assigned a business owner, and this owner should classify the data (for example, public, internal use only, or confidential). This provides assurance that the data is being protected in alignment with its sensitivity.
 - Determine the business owner of the data contained within the system and ask for evidence that the data has been classified according to your company's data classification system. This classification should appear on any reports or transactions that display system data. Also, determine whether the application's access control mechanisms are appropriate based on the classification.
 - Consider that data repositories may contain various kinds of data from different sources. Some systems, such as file servers or SharePoint, may provide access controls and partitioning in such a way that data of different classifications may be stored on the same system but with separate logical access lists. The overall controls of the environment should be commensurate with the highest classification (most sensitive) of data stored within or managed by the system.

DATA CLASSIFICATION SCHEMES

Classification Scheme



<https://www.archives.gov/cui>



PUBLIC

Data that may be freely disclosed to the public

Marketing Materials
Contact Information
Price Lists
etc



INTERNAL ONLY

Internal data not meant for public disclosure

Battlecards
Sales Playbooks
Organizational Charts
etc



CONFIDENTIAL

Sensitive data that if compromised could negatively affect operations

Contracts with Vendors
Employee Reviews
etc



RESTRICTED

Highly sensitive corporate data that if compromised could put the organization financial or legal risk

IP
Credit Card Information
Social Security Numbers
PHI

<https://edge.siriuscom.com/security/7-steps-to-effective-data-classification>

Checklist for Auditing Big Data and Data Repositories

- ☐ 1. Audit the OS-level controls relevant to the base operating system(s) included in the environment.
- ☐ 2. Verify that the application has appropriate password controls and other authentication controls as appropriate. Also, determine whether default application account passwords have been changed.
- ☐ 3. Ensure that the data classification of the environment is understood and review the data ownership process for the environment.
- ☐ 4. Review the system for the existence and use of role-based access controls and the processes for granting privileged access.
- ☐ 5. Review processes for granting and removing user access to view or search data. Ensure that access is granted only when there is a legitimate business need.
- ☐ 6. Ensure that company search systems follow data permissions rules if repositories or reports are indexed by systems outside of the repository scope.
- ☐ 7. Assess data retention, backup, and recovery procedures.
- ☐ 8. Review controls surrounding configuration management.
- ☐ 9. Review and evaluate procedures for monitoring and maintaining the security of the system.
- ☐ 10. Review governance processes for adding data sources to the big data environment.
- ☐ 11. Ensure that credentials or other methods used to load remote data sources are properly secured.