

Big Data Management and Security

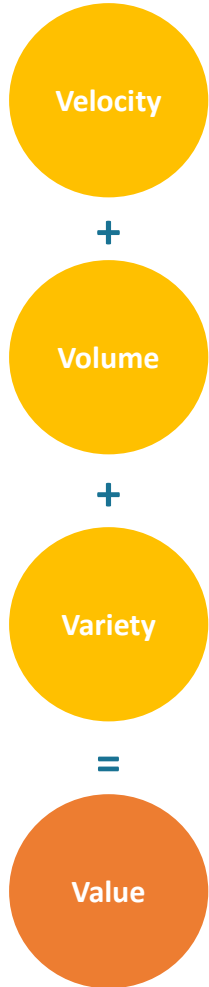
Audit Concerns and Business Risks

Tami Frankenfield

Sr. Director, Analytics and Enterprise Data


Mercury Insurance

What is Big Data?




Velocity

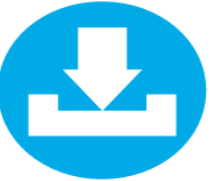
Frequency of data generation



48 hours
of video uploaded
to YouTube every
minute



2,000,000
queries
on Google every
minute




47,000
App downloads per
minute at the
Apple Store


Volume

The growth of world data

what is a
zettabyte

1,000,000,000,000	gigabytes
1,000,000,000	terabytes
1,000,000	petabytes
1,000	exabytes
1	zettabyte






1 terabyte holds the equivalent of roughly 210 single sided DVDs.


Variety

Structured and unstructured data – types of Big Data




Web and social media

Data includes clickstream and interaction data from social media such as Facebook, Twitter, LinkedIn, and blogs.




Machine to machine

Data includes readings from sensors, meters, and other devices as part of the so-called "internet of things."




Big transaction data

Includes healthcare claims, telecommunications call detail records (CDRs), and utility billing records that are increasingly available in semi-structured and unstructured formats.



Biometric

Data includes fingerprints, genetics, handwriting, retinal scans, and similar types of data.

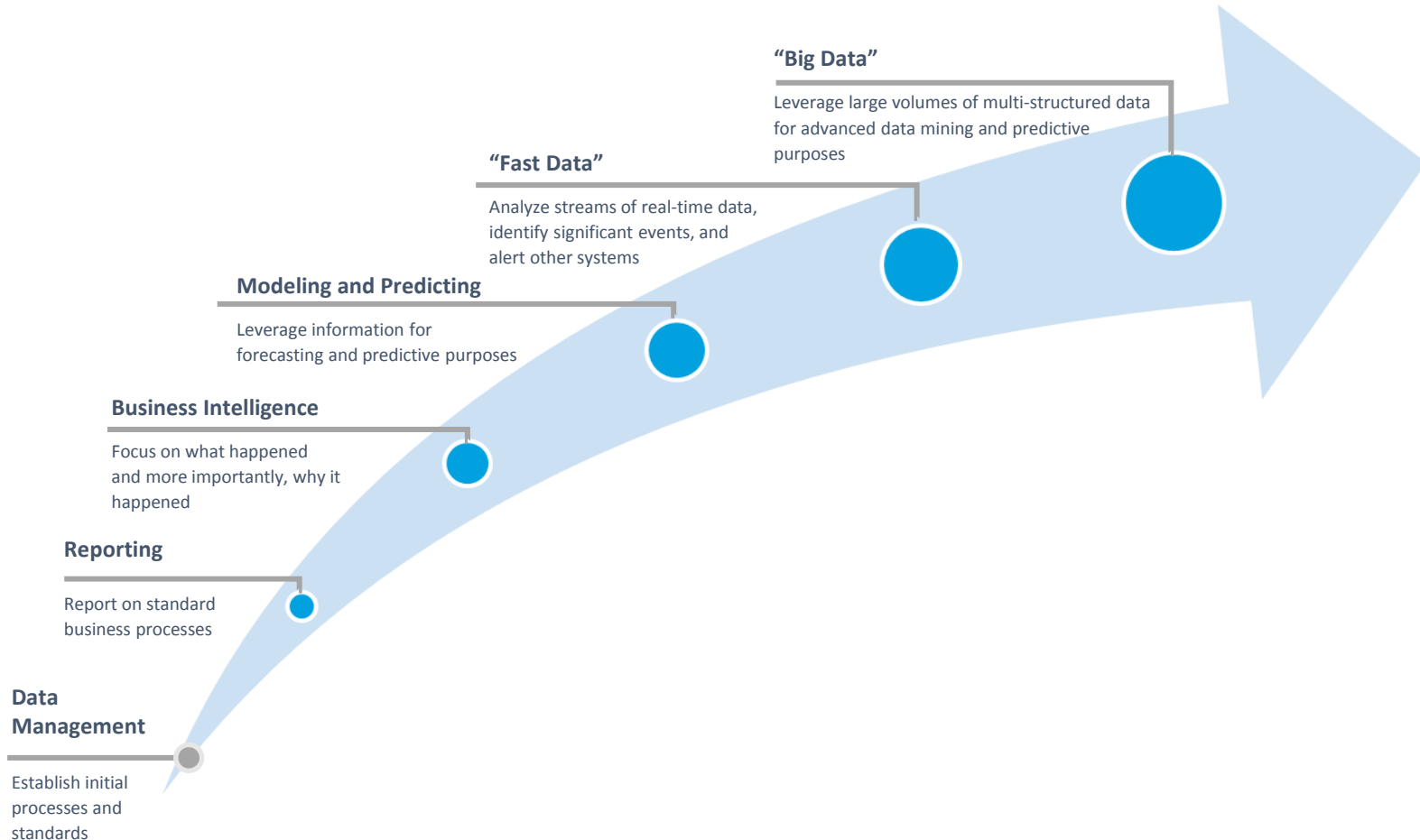


Human-generated

Data includes vast quantities of unstructured and semi-structured data such as call centre agents' notes, voice recordings, email, paper documents, surveys, and electronic medical records.

The Big Data Journey

Big Data is the next step in the evolution of analytics to answer critical and often highly complex business questions. However, that journey seldom starts with technology and requires a broad approach to realize the desired value.



Implications of Big Data?

Enterprises face the challenge and opportunity of storing and analyzing Big Data.

- Handling more than 10 TB of data
- Data with a changing structure or no structure at all
- Very high throughput systems: for example, in globally popular websites with millions of concurrent users and thousands of queries per second
- Business requirements that differ from the relational database model: for example, swapping ACID (Atomicity, Consistency, Isolation, Durability) for BASE (Basically Available, Soft State, Eventually Consistent)
- Processing of machine learning queries that are inefficient or impossible to express using SQL

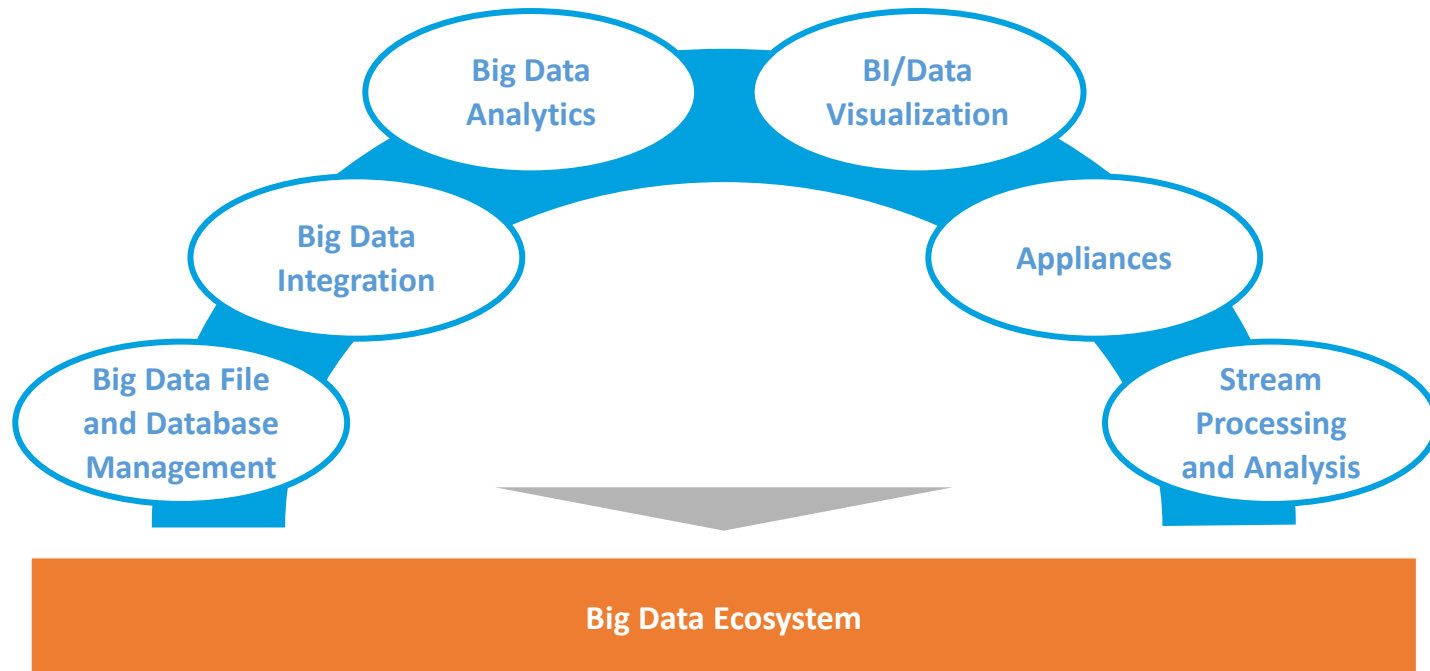
“Shift thinking from the old world where data was scarce, to a world where business leaders demonstrate data fluency” - Forrester

“Information governance focus needs to shift away from more concrete, black and white issues centered on ‘truth’, toward more fluid shades of gray centered on ‘trust.’ ” - Gartner

“Enterprises can leverage the data influx to glean new insights – Big Data represents a largely untapped source of customer, product, and market intelligence” – IBM CIO Study

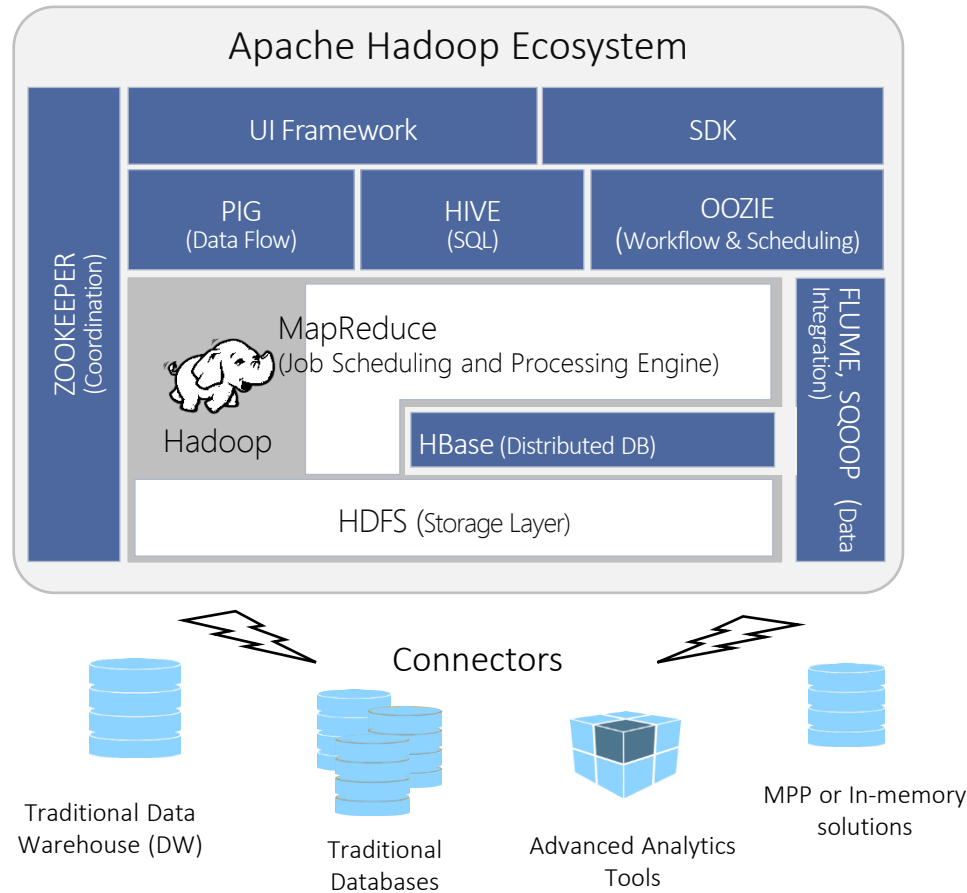
Taking a Look at the Big Data Ecosystem

Big Data is supported and moved forward by a number of capabilities throughout the ecosystem. In many cases, vendors and resources play multiple roles and are continuing to evolve their technologies and talent to meet the changing market demands.



Big Data Storage and Management

An Hadoop based solution is designed to leverage distributed storage and a parallel processing framework (MapReduce) for addressing the big data problem. Hadoop is an Apache foundation open source project.



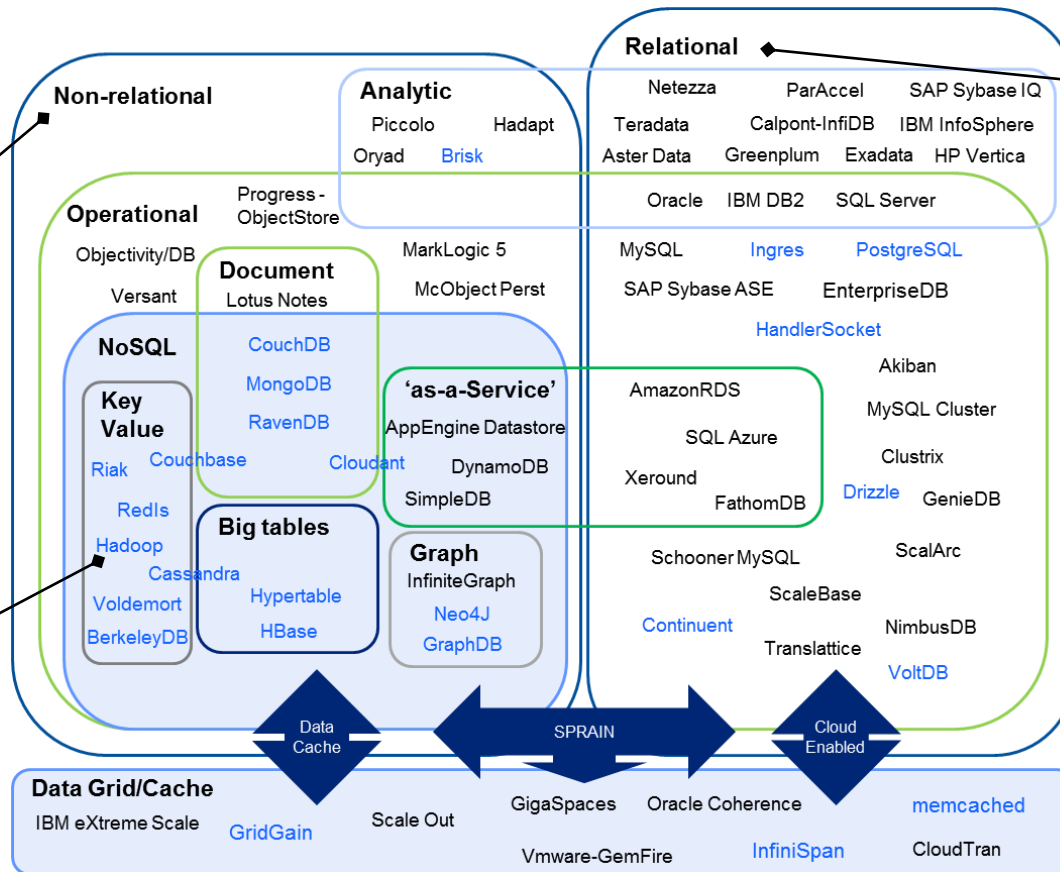
Big Data Storage and Management

The need for Big Data storage and management has resulted in a wide array of solutions spanning from advanced relational databases to non-relational databases and file systems. The choice of the solution is primarily dictated by the use case and the underlying data type.

Non-Relational Databases have been developed to address the need for semi-structured and unstructured data storage and management.

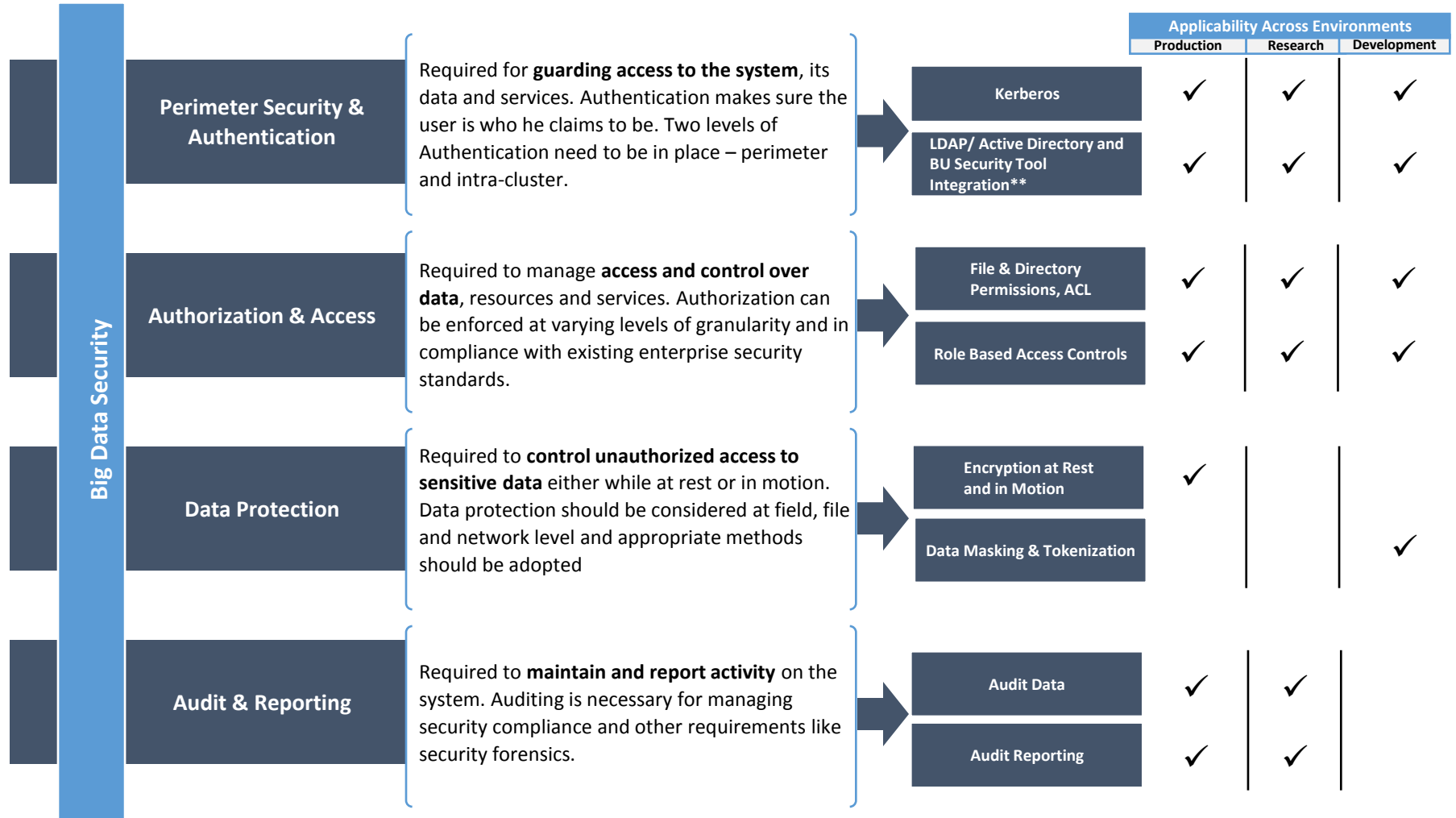
Hadoop HDFS is a widely used key-value store designed for Big Data processing.

Relational Databases are evolving to address the need for structured Big Data storage and management.



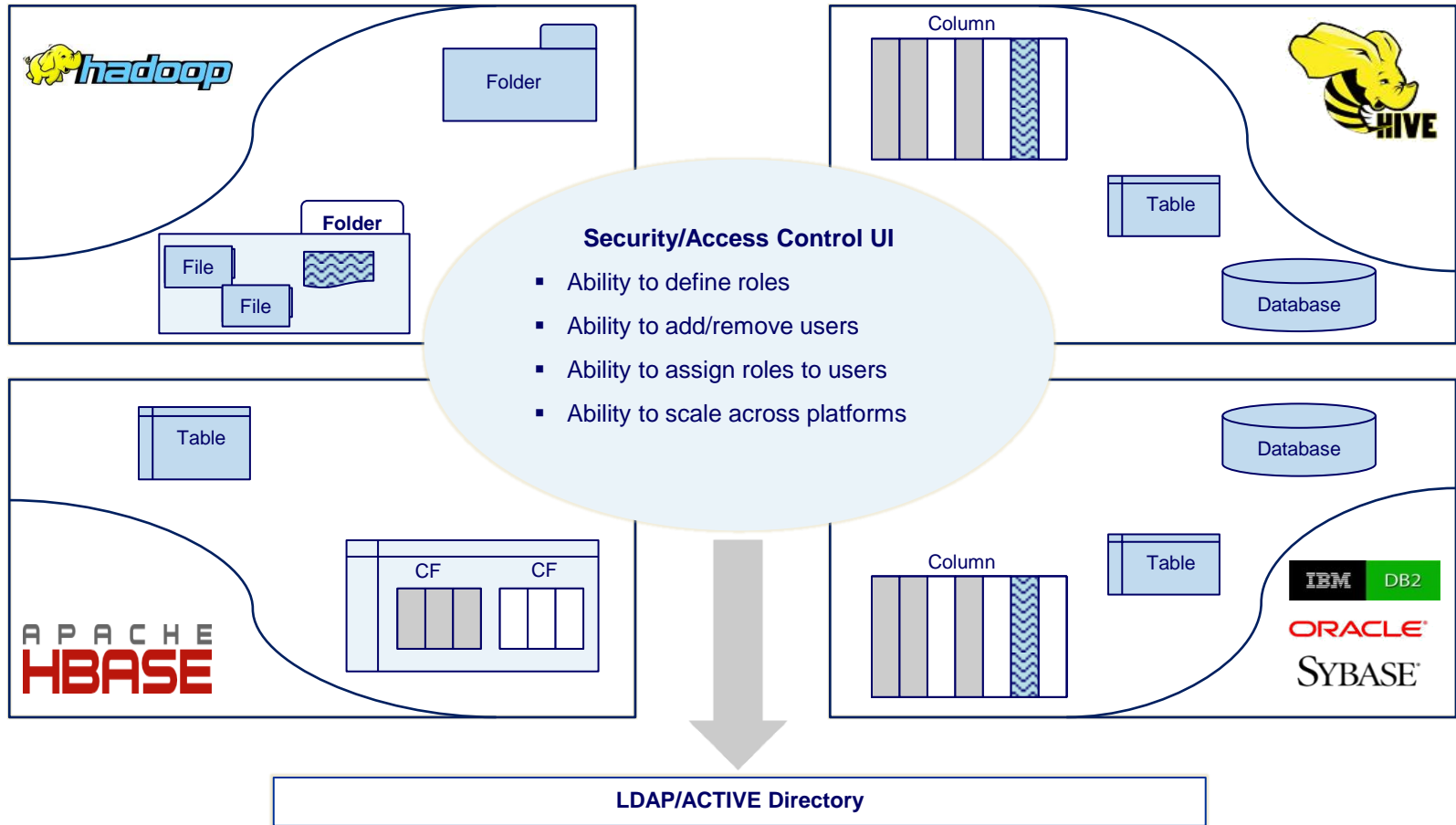
Big Data Security Scope

Big Data security should address four main requirements – perimeter security and authentication, authorization and access, data protection, and audit and reporting. Centralized administration and coordinated enforcement of security policies should be considered.



Big Data Security, Access, Control and Encryption

Integration of Security, Access, Control and Encryption across major components of the Big Data landscape.



Security, Access, Control and Encryption Details

Guidelines & Considerations

Encryption / Anonymization

- Data should be natively encrypted during ingestion of data into Hadoop (regardless of the data getting loaded into HDFS/Hive/HBase)
- Encryption Key management should be maintained at a Hadoop Admin level, there by the sanctity of the Encryption is maintained properly

Levels of granularity in relation to data access and security

- **HDFS:** Folder and File level access control
- **Hive:** Table and Column level access control
- **HBase:** Table and Column Family level access control

Security implementation protocol

- Security/Data Access controls should be maintained at the lowest level of details within a Hadoop cluster
- Overhead of having Security/Data Access should be minimum on any CRUD operation

Manageability / scalability

- GUI to Create/Maintain Roles/Users etc. to enable security or data access for all areas of Hadoop
- Ability to export/share the information held in the GUI across other applications/platforms
- Same GUI interface or application should be able to scale across multiple platforms (Hadoop and Non-Hadoop)

Key Terms Defined

In Hadoop, Kerberos currently provides two aspects of security:

- **Authentication** – This feature can be enabled by mapping the UNIX level Kerberos IDs to that of Hadoop. In a mature environment, Kerberos is linked/mapped to Active Directory or LDAP system of an organization. Maintenance of mapping is typically complicated
- **Authorization** – Mapping done in the Authentication level is leveraged by the Authorization and the users can be authorized to access data at the HDFS folder level

Point of View

Large organizations should adopt a more scalable solution with finer grains of access control and encryption / anonymization of data

Select a tool which is architecturally highly scalable and consists of the following features:

- Levels of granularity in relation to data access and security
- Security implementation protocol
- Manageability / scalability
- Encryption / Anonymization

Multi-Tenancy Details

Guidelines & Considerations

Multi-tenancy in Hadoop primarily need to address two key things:

1. Resource sharing between multiple applications and making sure none of the application impacts the cluster because of heavy usage
2. Data security/Auditing - users and applications of one application should not be able to access HDFS data of other apps
 - Vendors vary according to their support for a POSIX file system: MapR provides it by default, IBM BigInsights and Pivotal provide POSIX compliant add on packages (General Parallel File System and Isilon, respectively)
 - POSIX compliant file systems simplify initial migration into the Hadoop distribution. The relative advantage over other Hadoop distribution decreases as the load approaches steady state
 - Access Control Lists (ACLs) facilitate multi-tenancy by ensuring only certain groups and users can run jobs
 - With the evolution of YARN, the capacity scheduler can be used to set a minimum guaranteed resource per application as a % of RAM available (% of CPU will be possible in future)
 - This is a more efficient way of sharing resources between different groups within an organization. Before YARN, resources in Hadoop are available only as a number of map reduce slots available. So although multi-tenancy was possible, it was not very efficient

Key Terms Defined

- **A POSIX compliant filesystem** - provides high availability of Hadoop by having name node data totally distributed removing the single point of failure efficiently. Enables HDFS federation
- **MapR volume** – a collection of directories that contains data for a single business unit, application or user group. Policies may be applied to a volume to enforce security or ensure availability
- **YARN** decouples workload management from resource management, enabling multi tenancy

Point of View

- Multi-tenancy begins with a multi-user-capable platform. Further, in a clustered computing environment, it involves multi-tenancy at a data (file system) level, a workload (jobs and tasks) level, and a systems (node and host) level
- The recent advancements of YARN and Hadoop 2.0 are quickly closing the feature gap between proprietary file systems and HDFS

Big Data Security Options and Recommendations

	Background	Recommendations
Perimeter Security & Authentication	<ul style="list-style-type: none">▪ Hadoop supports Kerberos as a third party Authentication mechanism▪ Used for Intra Services authentication (Name Node-Data Node, Task Tracker, Oozie etc),End User-Services (Hue, File browser, cli etc)▪ Ticket based authentication and Operates within a realm(hadoop cluster) and inter-realms as well▪ Users and services rely on a third party Kerberos server - to authenticate each other▪ Offers one way, two way Trust options▪ Offers LDAP integration options	<ul style="list-style-type: none">▪ Kerberos is highly recommended as it supports authentication mechanisms throughout the cluster▪ Use Apache Knox for perimeter authentication to Hive, HDFS, HBase etc▪ Manage user groups in POSIX layer (initially atleast). Resolving user groups from LDAP involves corporate IT dependency▪ Hadoop security implementation is not easy. Take a phased approach.<ul style="list-style-type: none">✓ Configure Kerberos for Hadoop✓ Provision initial set of users/user groups in POSIX layer✓ Integrate LDAP/Single-Sign On (Kerberos, Hue, Ambari)✓ Prepare final list of users/user groups and provision the users on POSIX layer matching ldap principals✓ Optionally Configure Knox gateway for perimeter authentication for external systems (with LDAP or SSO)
Authorization & Access	<ul style="list-style-type: none">▪ HDFS Permission bits▪ HDFS ACLs▪ YARN ACLs▪ Access control for Hive/HBase with Apache Knox	<ul style="list-style-type: none">▪ Authorization control with Apache Knox for column/row access restrictions for users▪ Optionally configure Accumulo if cell level restrictions are required for HBase/Hive
Data Protection	<ul style="list-style-type: none">▪ Hadoop supports Data Encryption in Motion – HTTPS(web clients, REST), RPC Encryption (API, Java Frameworks), Data Transfer protocol (Name Node-Data Node)▪ No Options available for Data Encryption at Rest at the moment	<ul style="list-style-type: none">▪ Configure Data in Motion wire encryption▪ Agree on data encryption algorithms for data which needs to be exported outside lake within Zurich network▪ Optionally implement 'Data at Rest' Encryption
Audit & Reporting	<ul style="list-style-type: none">▪ HDFS Auditing – Centrally available via Ranger(XA secure)▪ Perimeter Access – Available at Knox▪ Job Auditing – Admin console, Job Tracker Logs	<ul style="list-style-type: none">▪ Monitor your YARN capacity – Key to enhance multi-tenancy▪ Set-up Security compliance process▪ Set-up user administrative process/auditing