Unit 2: MarkLogic Server Architecture

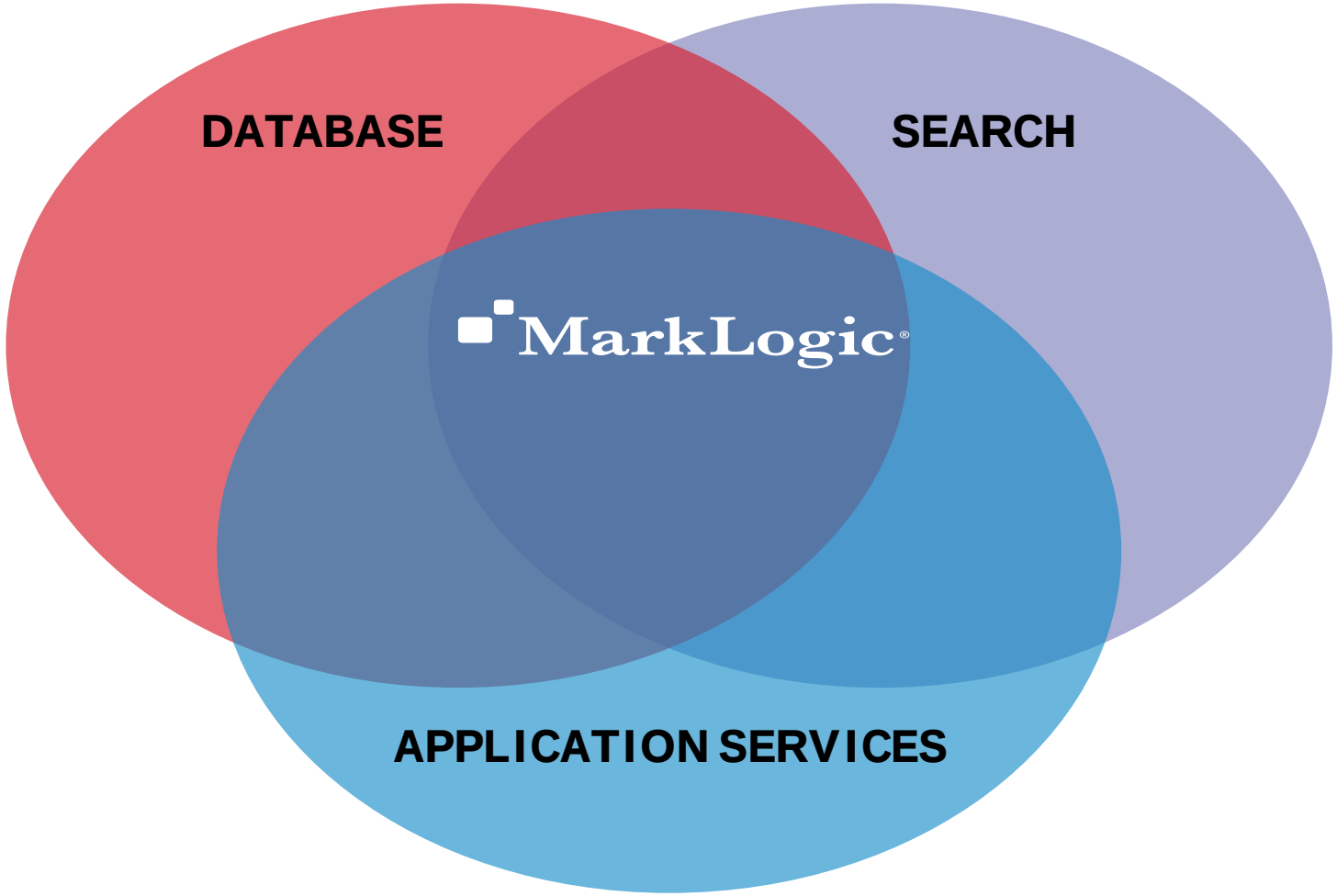
MarkLogic

Learning Objectives

- Describe each of the following:
  - Database (Content / Modules / Schema / Triggers / Security)
  - Forest
  - Stand (In Memory / On Disk)
  - Universal Index
  - Application Server (HTTP / XDBC / ODBC / WebDAV)
  - Evaluator Node | Database Node
  - Hosts | Groups
  - List Cache | Compressed Tree Cache | Expanded Tree Cache
- Describe storage options in MarkLogic
- Compare / contrast a single host vs. cluster implementation

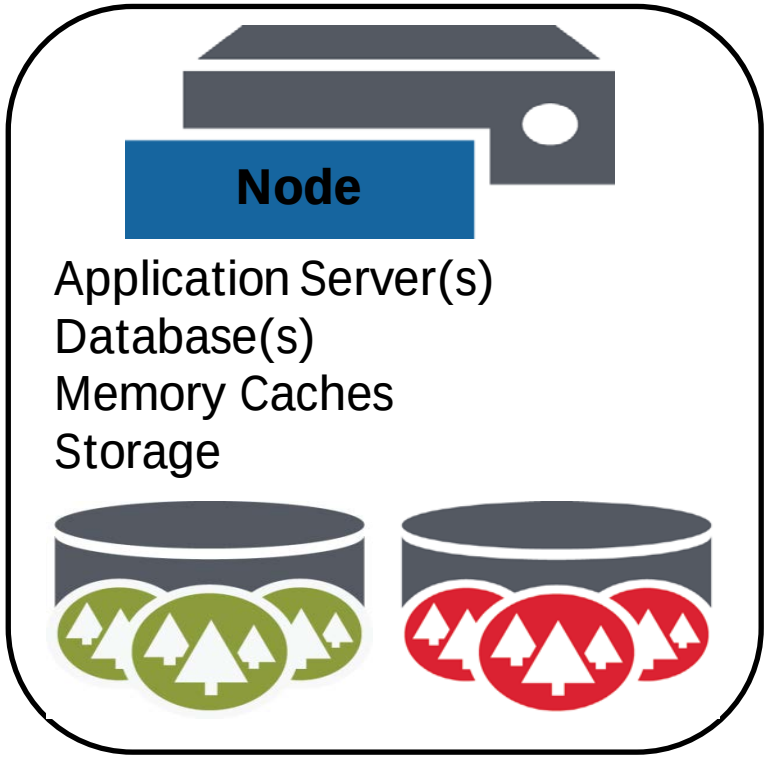
MarkLogic

MarkLogic Server

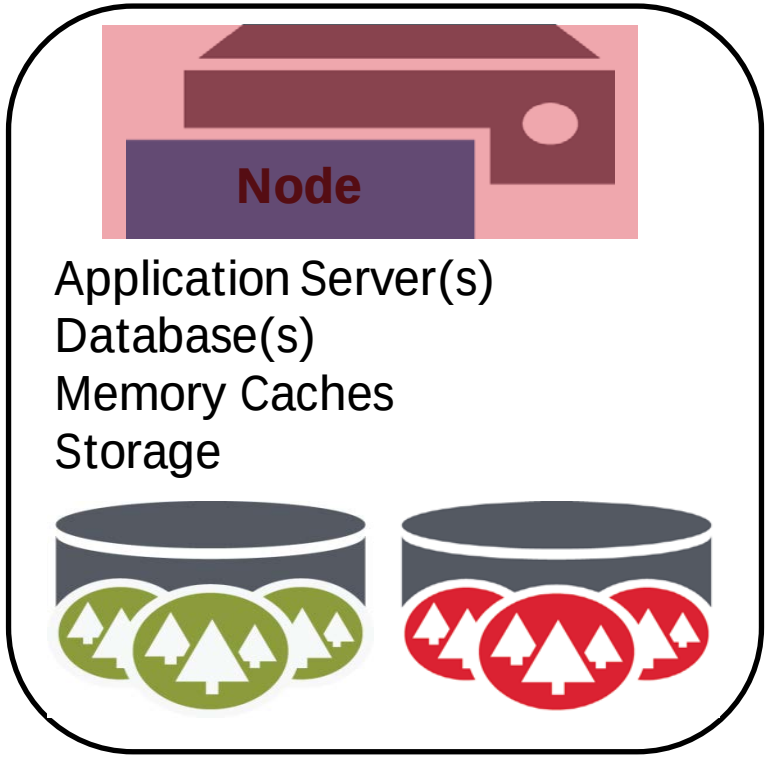


Single Node Architecture

- Example: Our training environment
- All MarkLogic resources configured on one machine
  - Node = Host = a machine running MarkLogic

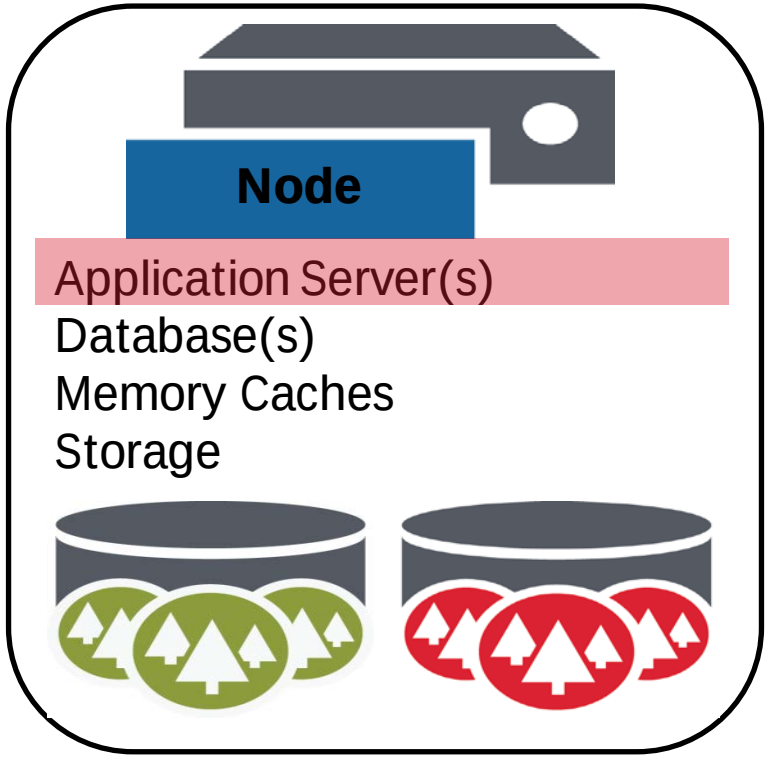


Single Node Architecture

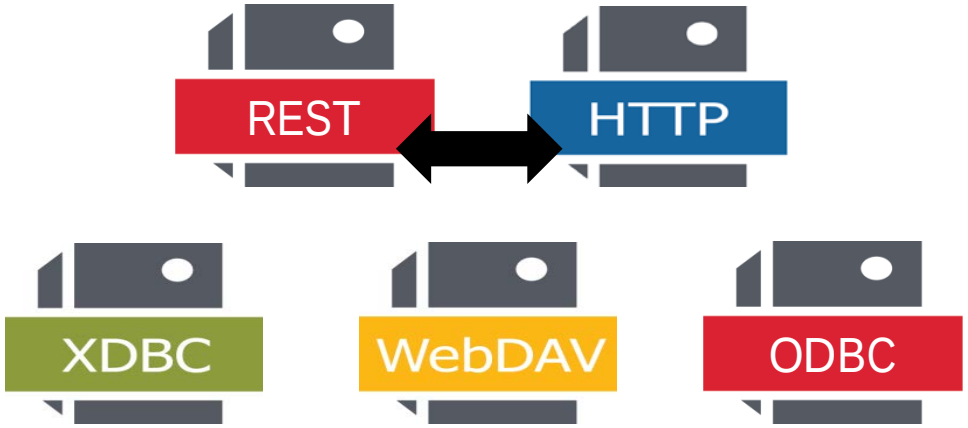


- Node / Host / Machine
  - E = Evaluator = App Server
  - D = Data Manager = Database
- MarkLogic runs as a service
- A machine can act as both (E/D)
- In a cluster, machines may specialize

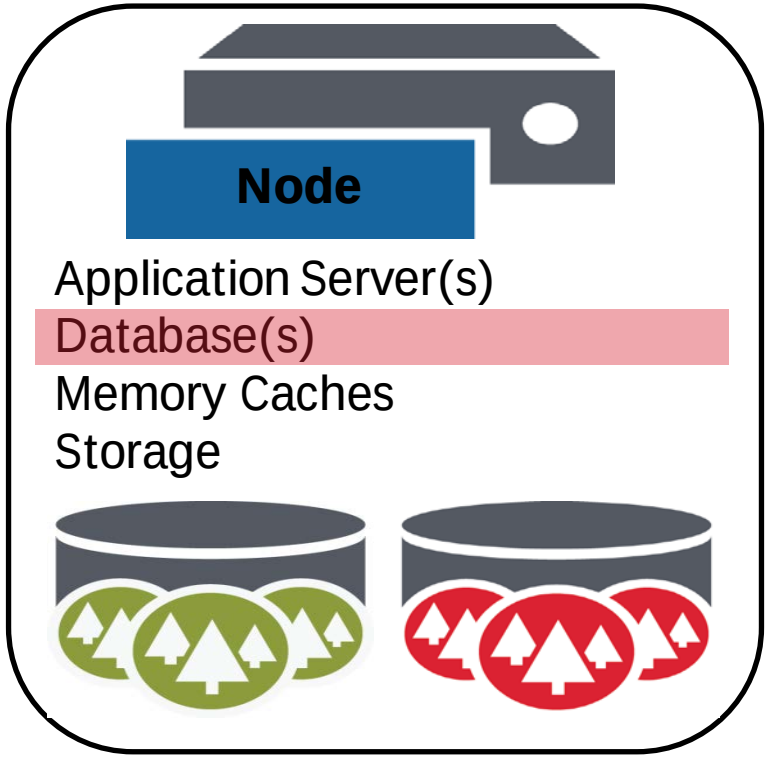
Single Node Architecture



- Application Server
  - Handles requests / responses
  - Defined on a port
  - Evaluates code

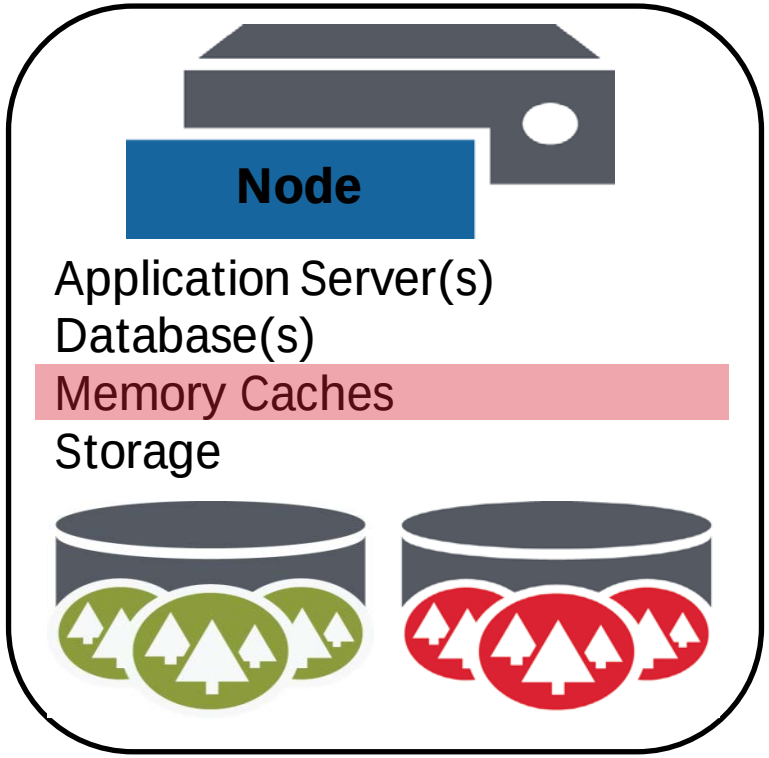


Single Node Architecture



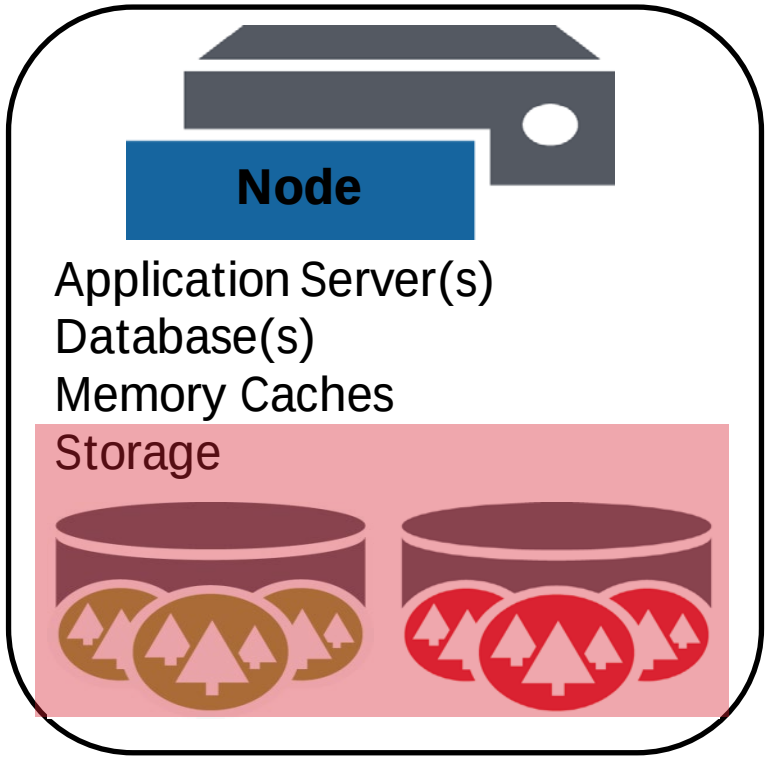
- Database
  - Transaction controller
  - Logical configuration
    - Indexing / Reindexing settings

Single Node Architecture



- Configurable memory caches:
  - List Cache
    - Recently accessed indexes
  - Compressed Tree Cache
    - Recently accessed docs, compressed
  - Expanded Tree Cache
    - Recently accessed docs, uncompressed
  - Triple Data and Triple Value Caches
    - Recently accessed triples

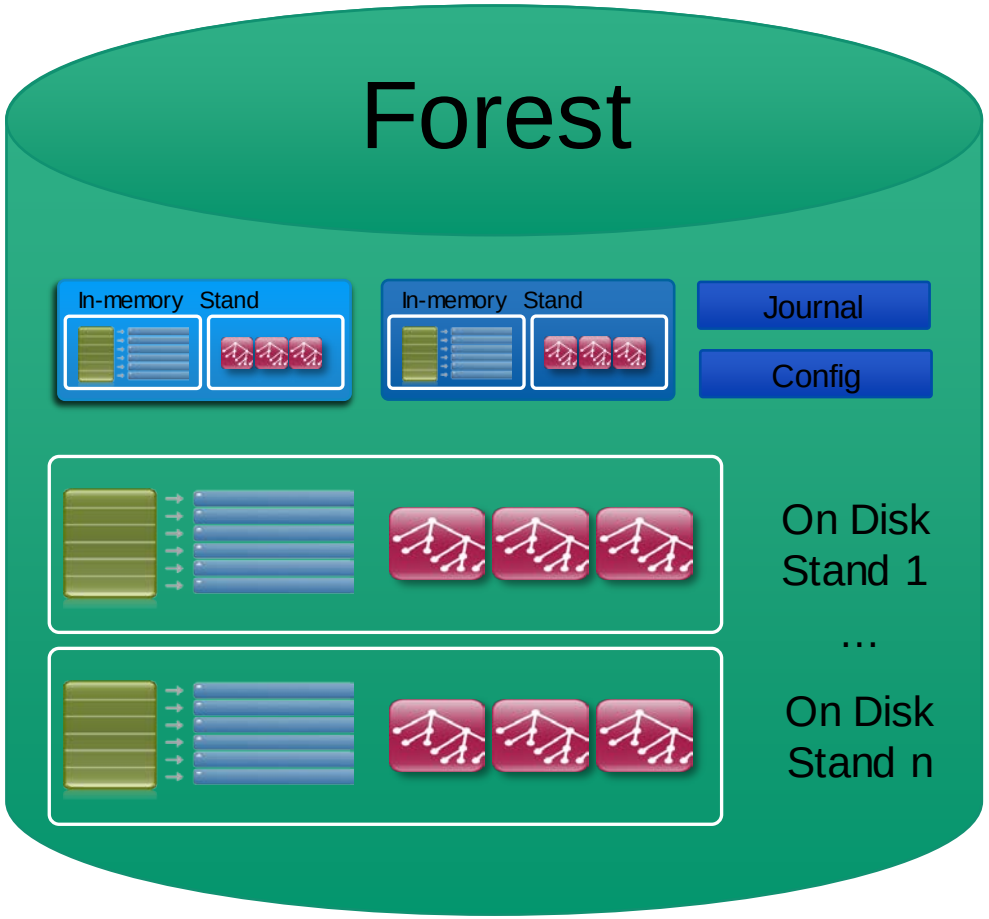
Single Node Architecture



- Forests
  - Physical storage
  - Attached to database (1DB:Many Forests)
  - Stands
    - Memory
    - Disk
    - Documents
    - Indexes
    - Compression
  - Journal

MarkLogic

Forest Details

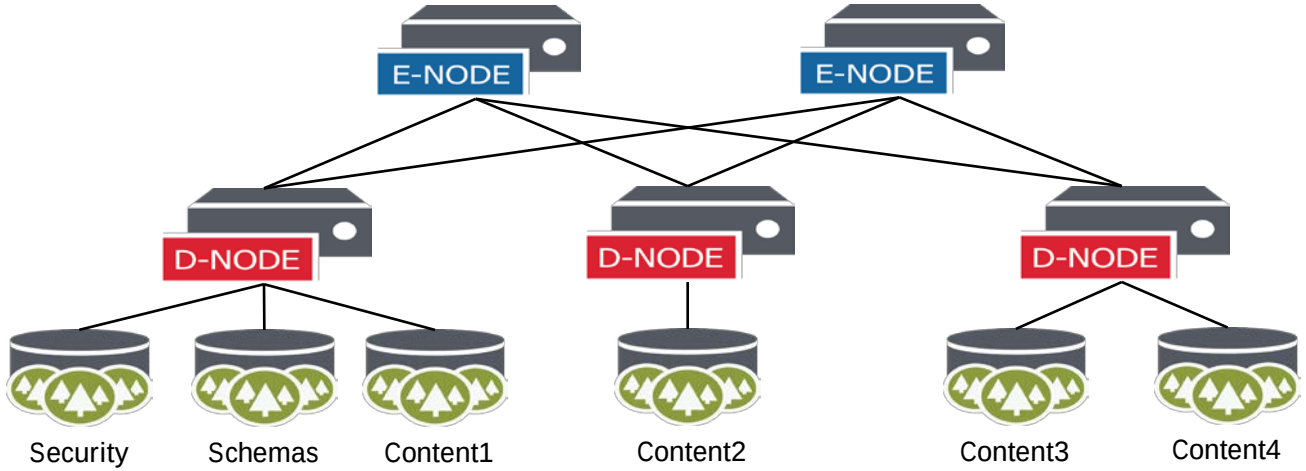


Demo:
Databases, Forests, Stands, Merges

MarkLogic

Clustered Architecture

- MarkLogic is a shared-nothing distributed database, allowing linear scale out and high availability
  - 3 node minimum for High Availability

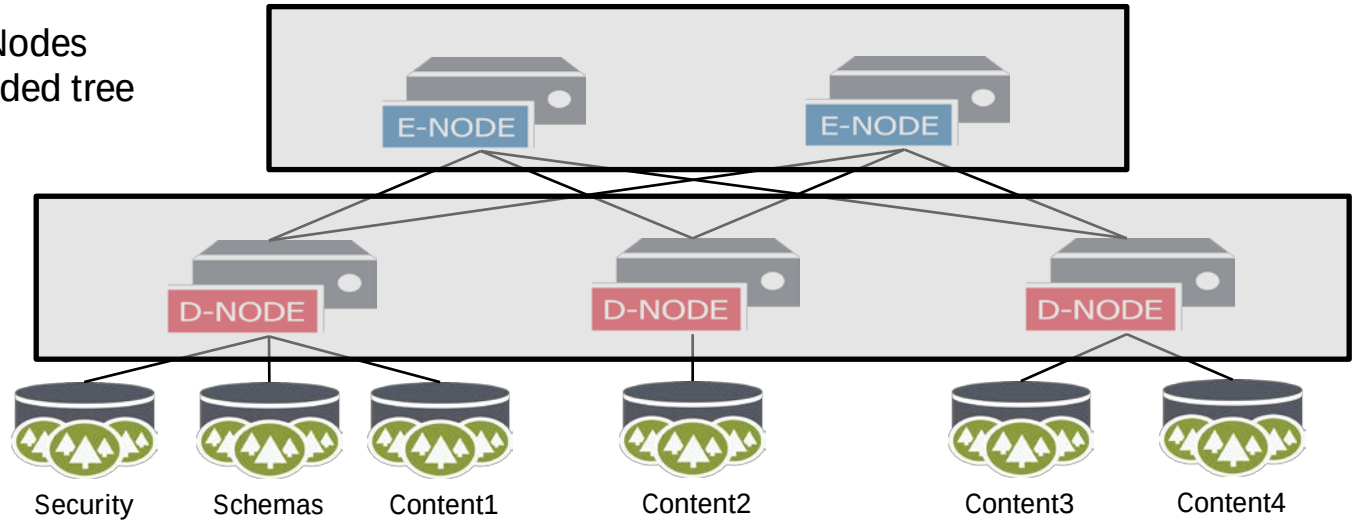


Clustered Architecture

- Groups
  - Groups of host machines within a cluster
  - Enables more specific configuration

Group 1: E Nodes
Bigger expanded tree cache

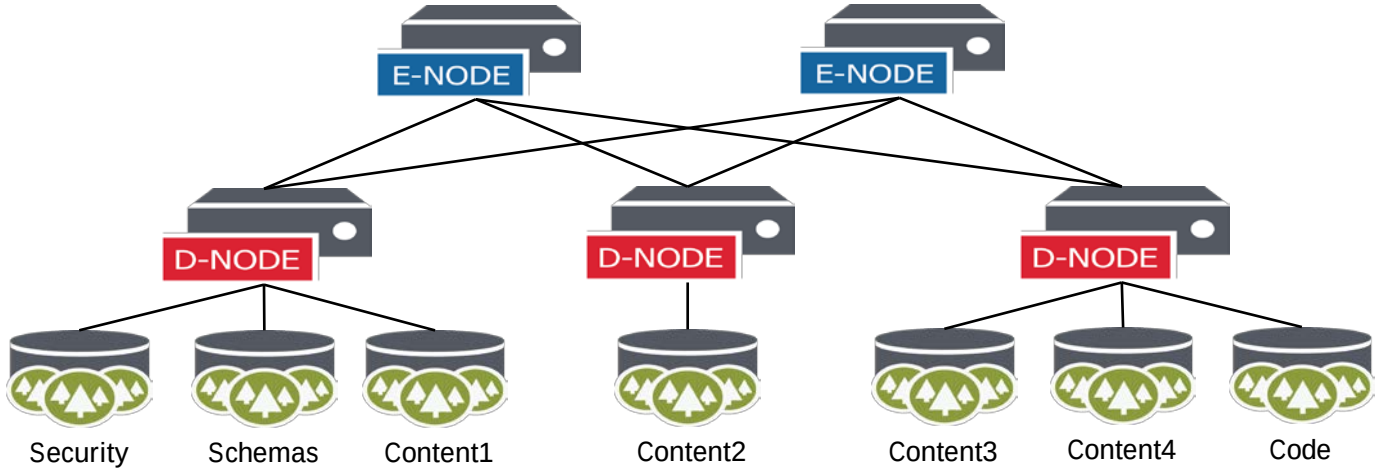
Group 2: D Nodes
Bigger compressed tree & list caches



MarkLogic

Clustered Architecture

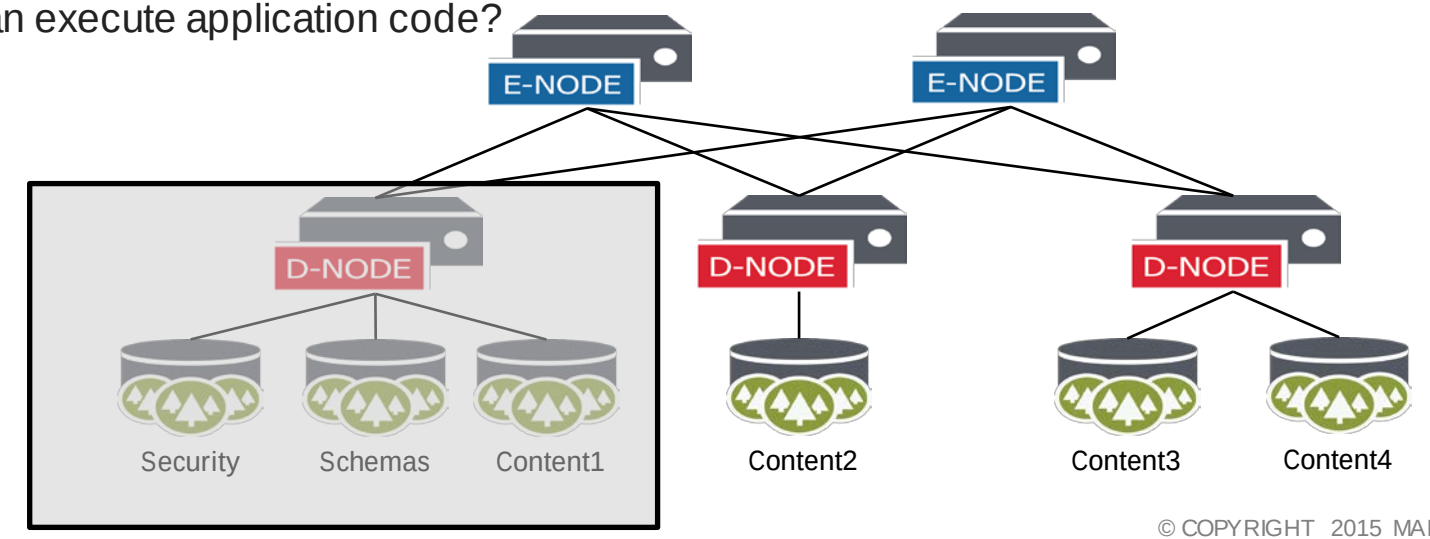
- Common Databases
  - Can be shared across multiple projects
  - Security | Schemas | Triggers | Modules
  - Impacts to HA / DR design



MarkLogic

Security 101

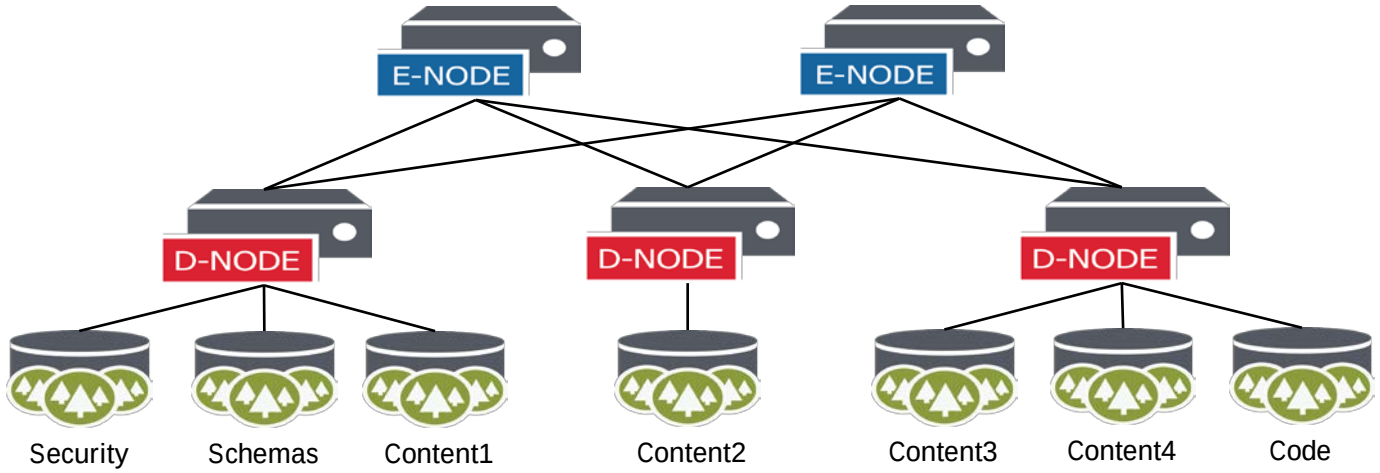
- Role based model
- Authentication
  - Performed on the application server
  - Utilize LDAP / Kerberos external authentication protocol
- Database Level
  - Who can read / write / update documents within a database?
- Code Level
  - Who can execute application code?



MarkLogic

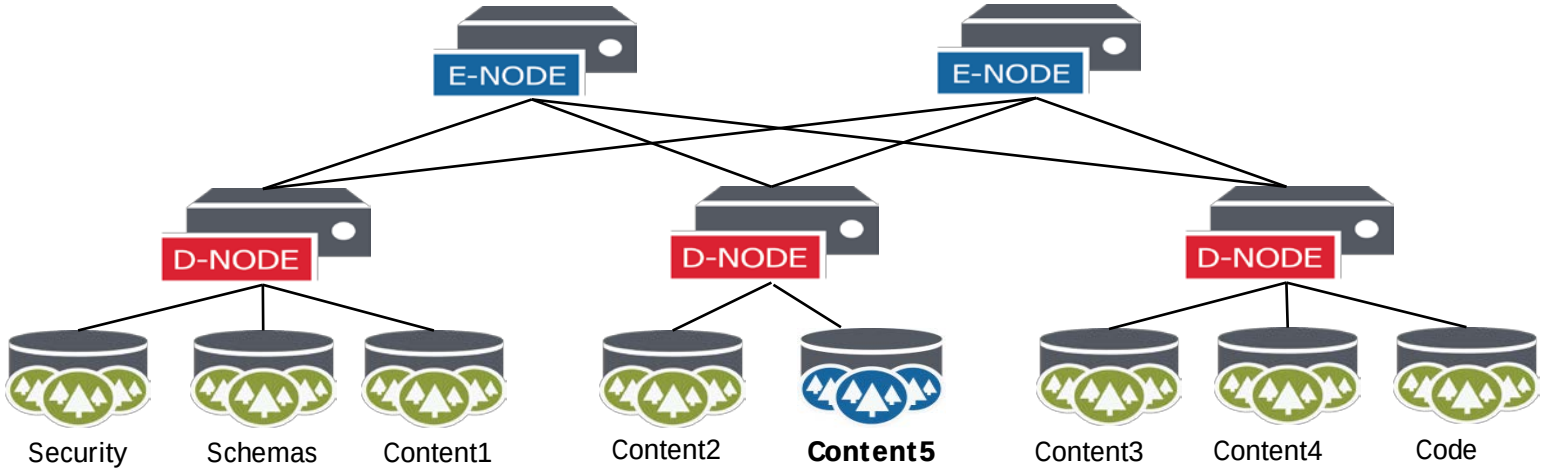
Clustered Architecture

- Scalability
  - Scaling a cluster to support more data and/or users
    - Add more forests
    - Add more E nodes and / or D nodes



Clustered Architecture

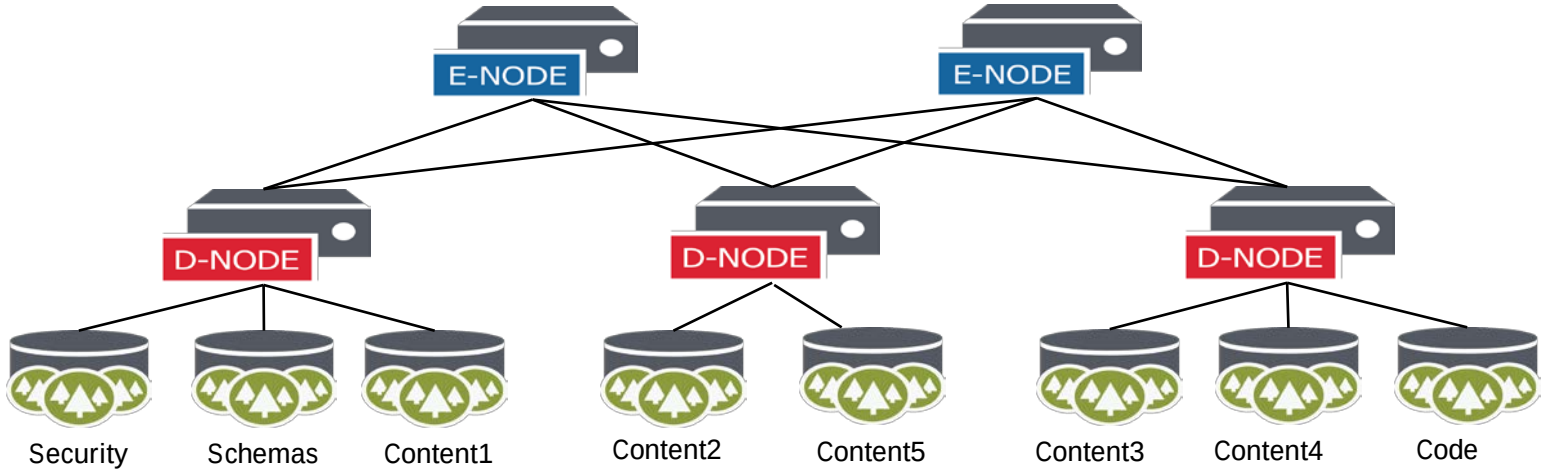
- Scalability & Rebalancing
  - Assume your data needs have increased
  - One of your D Nodes still has capacity
  - You add another forest – what happens when you load more data?
    - Pre-MarkLogic 7 (no automatic rebalancing)
    - MarkLogic 7 (rebalancing on by default)



MarkLogic

Clustered Architecture

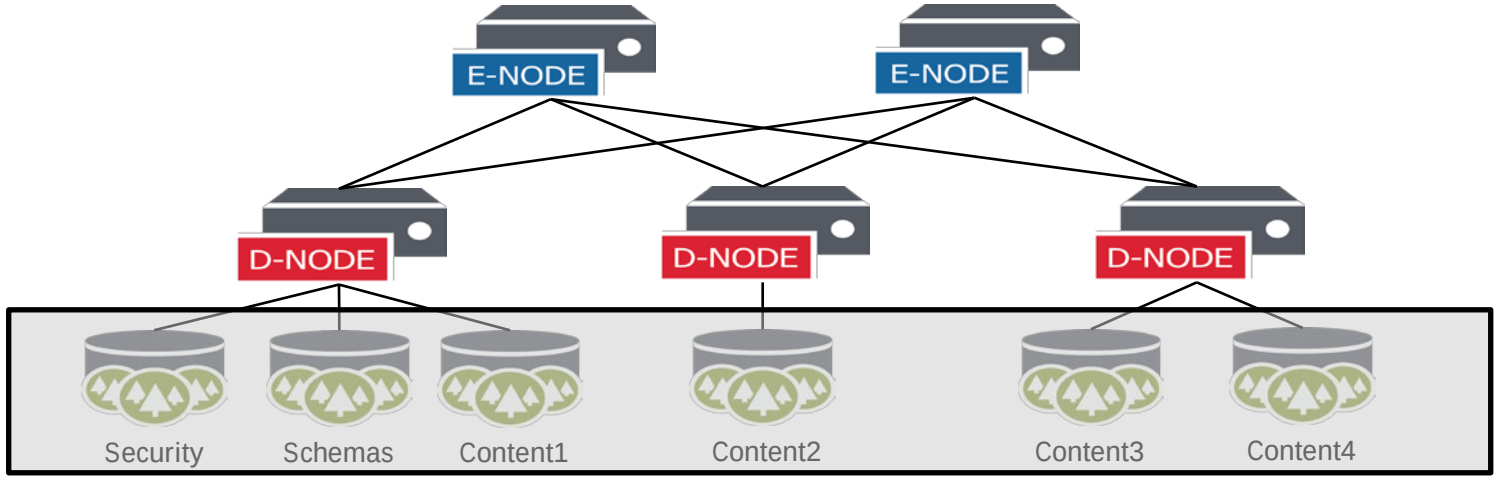
- Rebalancing
  - Administrator can choose appropriate **assignment policy** and the **rebalancer** will automatically handle the data movement



MarkLogic

Clustered Architecture

- Storage
  - Local, Shared, SSD, HDFS, Amazon S3
  - Tiered Storage
  - Optimized storage of large binary data



MarkLogic

More on storage…

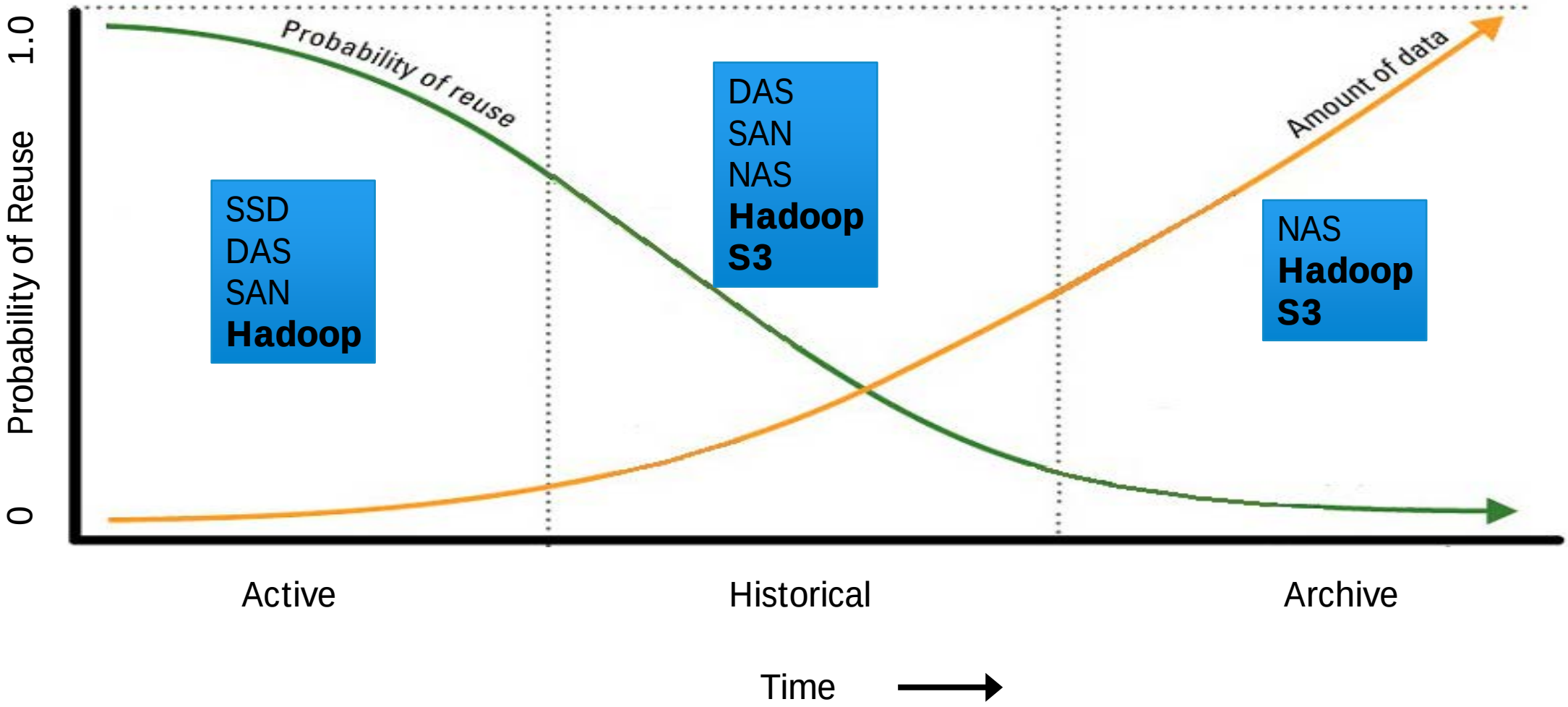
- MarkLogic enables you to choose a portfolio of storage options
- Design for your business, budget and performance goals

Rotating Disks	SSD	Amazon S3*	HDFS
Low Cost	Greater Cost	Low Cost	Low Cost
Performance impacted by many variables (controllers, latency, disk quality)	Fastest	Cloud based, globally distributed, access via HTTP, tight EC2 integration	Distributed, sequential I/O, configurable, replication
		Good for backups	

- *MarkLogic does not create journals on Amazon S3

MarkLogic

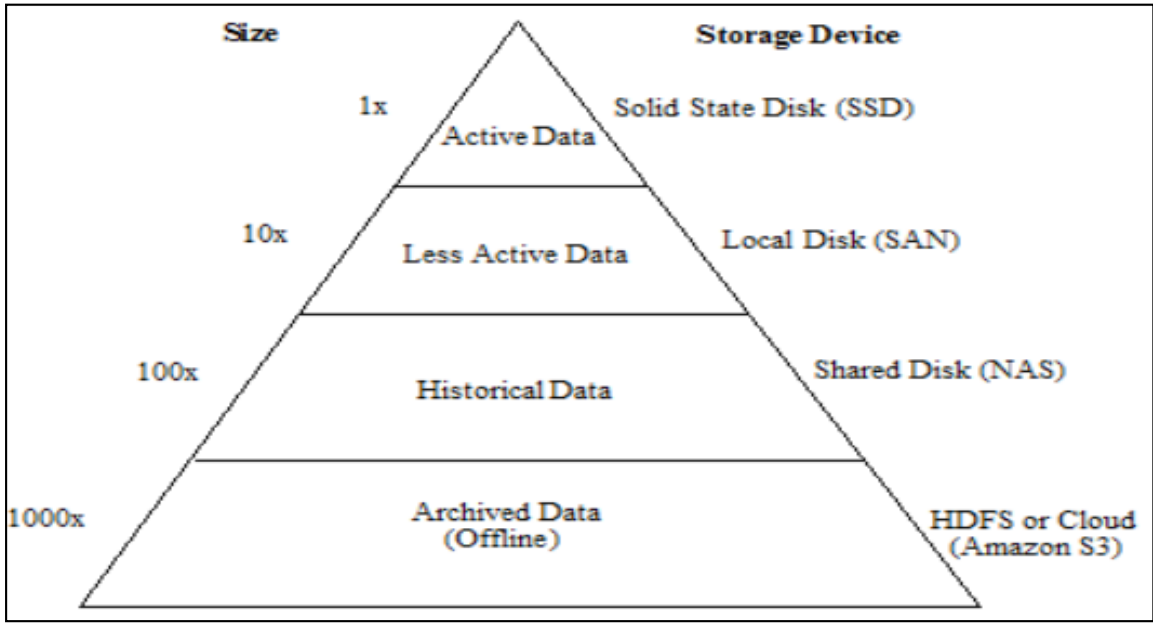
Storage and the Information Lifecycle



MarkLogic

Tiered Storage Overview

- Objective:
  - Manage data using a portfolio of storage options
  - Design for your performance objectives and cost constraints
  - Partitions = groups of forests
    - APIs enabling you to migrate, resize, move offline/online, delete



MarkLogic

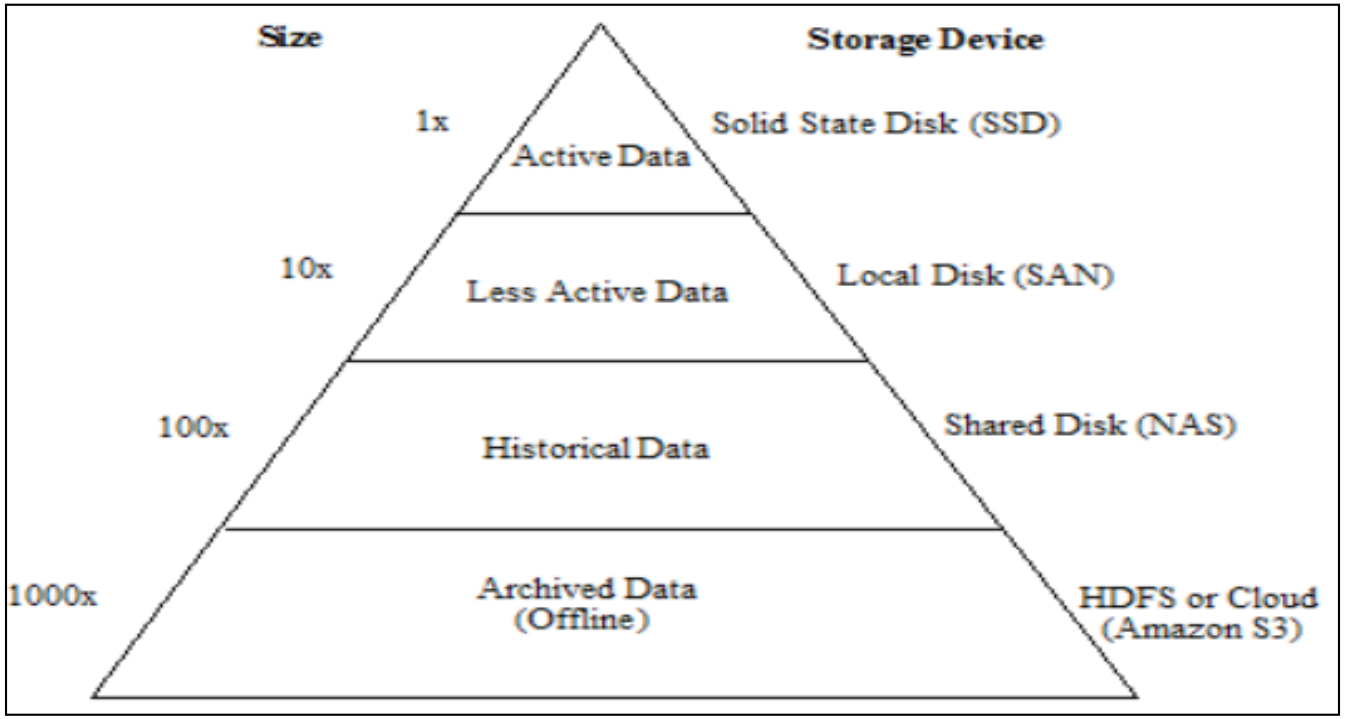
Tiered Storage Overview

- Conceptual implementation – where would these doc's end up?

```
<doc>
abc 123 xyz
<d>2013-10-22</d>
abc 123 xyz
</doc>
```

```
<doc>
abc 123 xyz
<d>1968-05-10</d>
abc 123 xyz
</doc>
```

```
<doc>
abc 123 xyz
<d>1985-08-16</d>
abc 123 xyz
</doc>
```



Partition Name	Forest Name	Range Definition <d>
V1	V1-01, V1-02	> 2013-01-01 < 2015-12-31 *Include LB
V2	V2-01, V2-02, V2-03	> 2000-01-01 < 2012-12-31 *Include LB
V3	V3-01, V3-02, V3-03, V3-04	> 1980-01-01 < 1999-12-31 *Include LB
V4	V4-01, V4-02, V4-03, V4-04, V4-05	> 1900-01-01 < 1979-12-31 *Include LB

MarkLogic

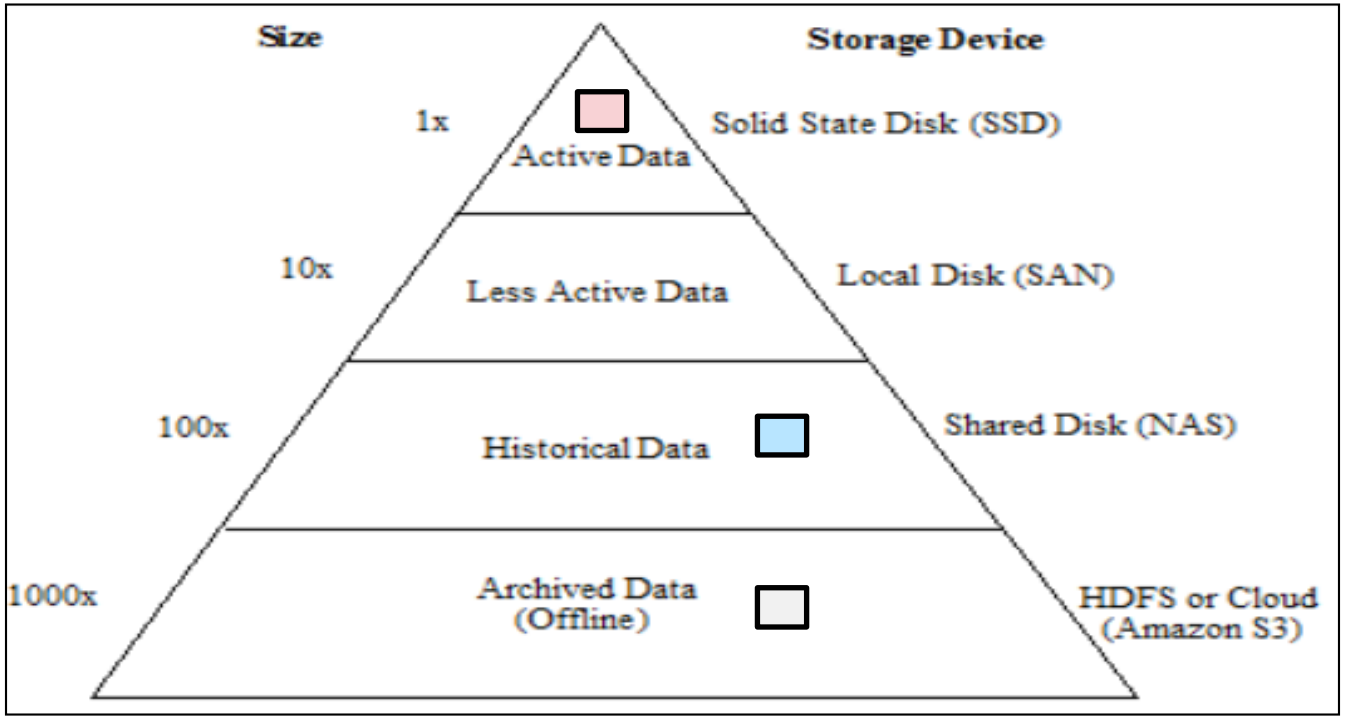
Tiered Storage Overview

- Conceptual implementation – where would these doc's end up?

<doc>
abc 123 xyz
<d>2013-10-22</d>
abc 123 xyz
</doc>

<doc>
abc 123 xyz
<d>1968-05-10</d>
abc 123 xyz
</doc>

<doc>
abc 123 xyz
<d>1985-08-16</d>
abc 123 xyz
</doc>

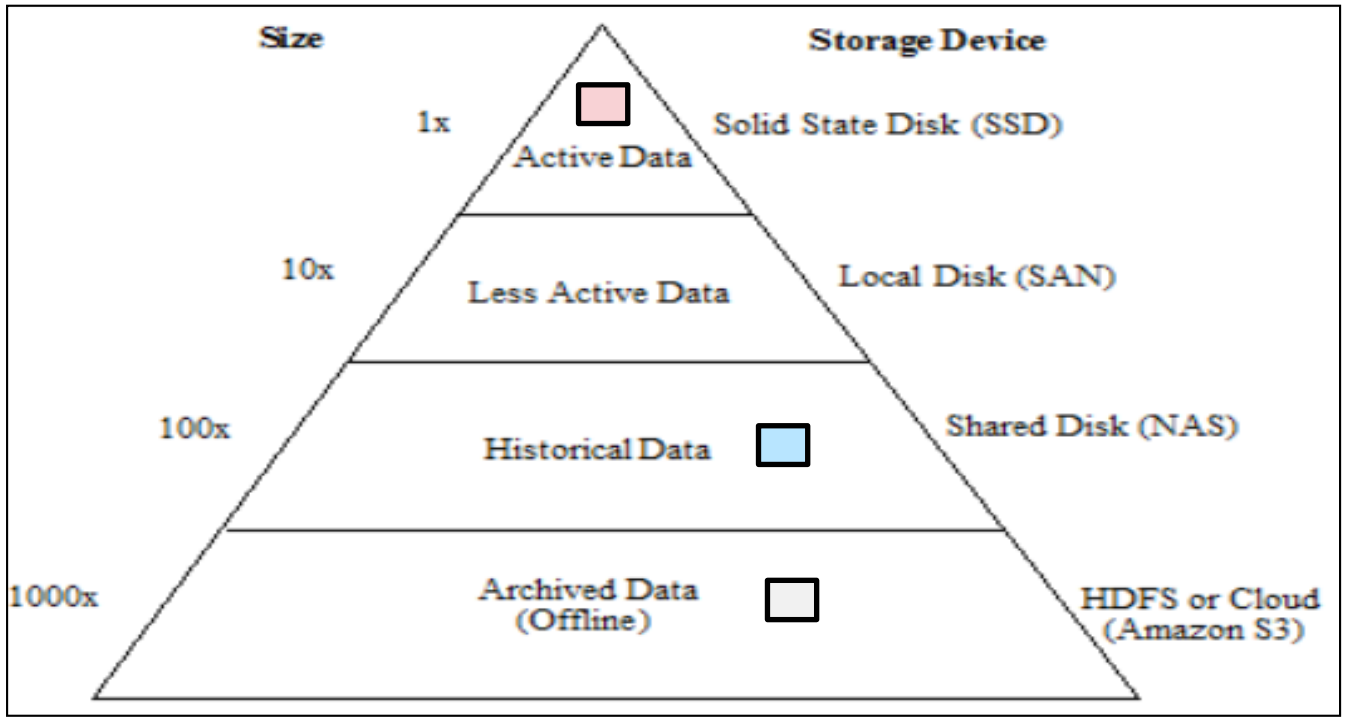


Partition Name	Forest Name	Range Definition <d>
V1	V1-01, V1-02	> 2013-01-01 < 2015-12-31 *Include LB
V2	V2-01, V2-02, V2-03	> 2000-01-01 < 2012-12-31 *Include LB
V3	V3-01, V3-02, V3-03, V3-04	> 1980-01-01 < 1999-12-31 *Include LB
V4	V4-01, V4-02, V4-03, V4-04, V4-05	> 1900-01-01 < 1979-12-31 *Include LB

MarkLogic

Tiered Storage Overview

- What about access?
  - Each tier can be accessed individually
  - Or combined into a single unified system



MarkLogic

MarkLogic Architecture Summary

Demo:
MarkLogic Architecture in Samplestack

Labs: Unit 2

Exercise 1: Explore Inside MarkLogic Server

Unit Review Question 1:

Which of the following is stored in a forest stand:

1. Database configuration information
2. Uncompressed data
3. Compressed data
4. Indexes

Unit Review Question 1:

Which of the following is stored in a forest stand:

1. Database configuration information
2. Uncompressed data
3. **Compressed data**
4. **Indexes**

Unit Review Question 2:

An instance of a database can only exist on one host in a cluster:

1. True
2. False

Unit Review Question 2:

An instance of a database can only exist on one host in a cluster:

1. True
2. **False**

Unit Review Question 3:

You run a search query against a MarkLogic database in a cluster to find a relevant document, read it from the database, and display the document content to an end user.

What will the following caches contain?

1. List
2. Compressed tree
3. Expanded tree
4. Triple

Unit Review Question 3:

You run a search query against a MarkLogic database in a cluster to find a relevant document, read it from the database, and display the document content to an end user.

What will the following caches contain?

1. List – indexes that were used in the search
2. Compressed tree – the compressed document
3. Expanded tree – the uncompressed document
4. Triple – nothing, assuming no SPARQL in the search

Unit Review Question 4:

You have added memory to your cluster and you want to increase the size of one of the MarkLogic caches.

At what level would this be configured?

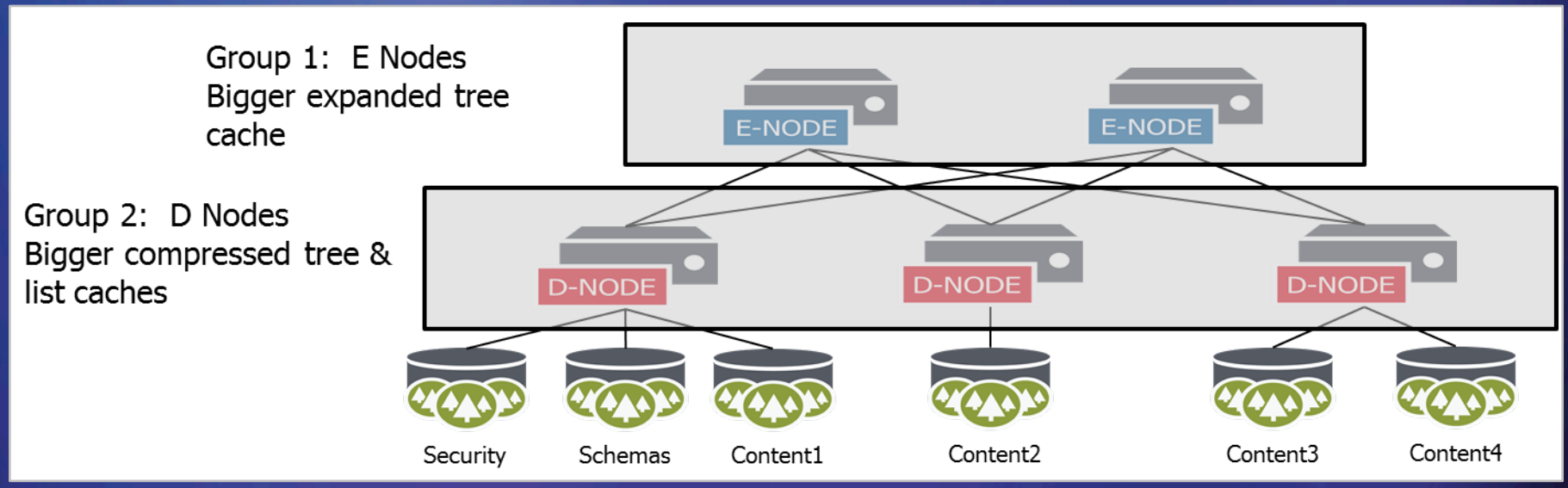
1. Host
2. App Server
3. Group
4. Cluster

Unit Review Question 4:

You have added memory to your cluster and you want to increase the size of one of the MarkLogic caches.

At what level would this be configured?

1. Host
2. App Server
3. **Group**
4. Cluster



Unit Review Question 5:

Assume you are building a cluster with a group of E Nodes and a group of D Nodes.

When building the cluster, the first host that you install and initialize should be an:

1. E Node
2. D Node
3. It doesn’t matter

Unit Review Question 5:

Assume you are building a cluster with a group of E Nodes and a group of D Nodes.

When building the cluster, the first host that you install and initialize should be an:

1. E Node
2. **D Node**
3. It doesn’t matter

Unit Review Question 6:

In a 3 tier architecture, where does the MarkLogic cluster fit?

1. Browser tier
2. App server / middle tier
3. Database tier

Unit Review Question 6:

In a 3 tier architecture, where does the MarkLogic cluster fit?

1. Browser tier
2. App server / middle tier
3. **Database tier**