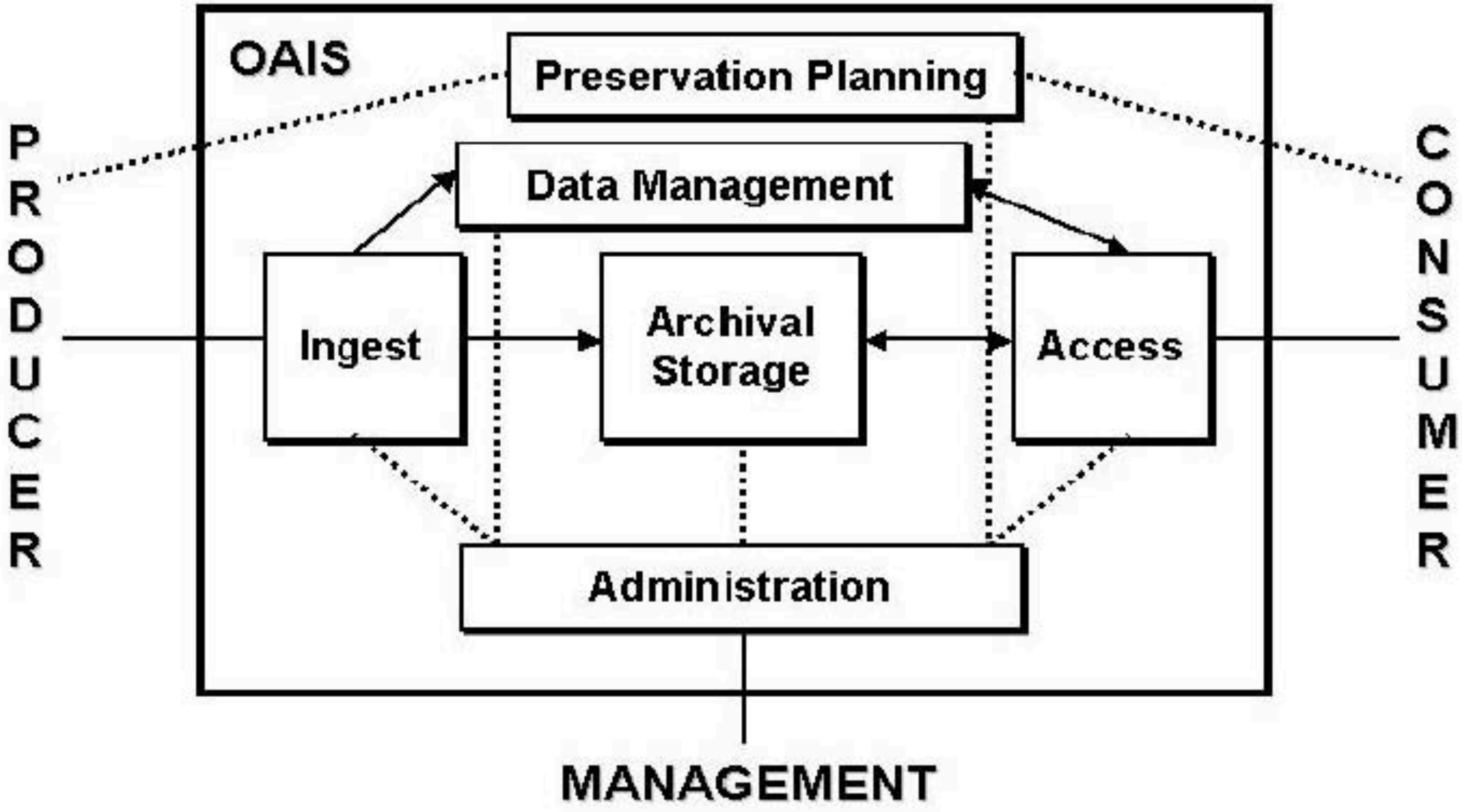
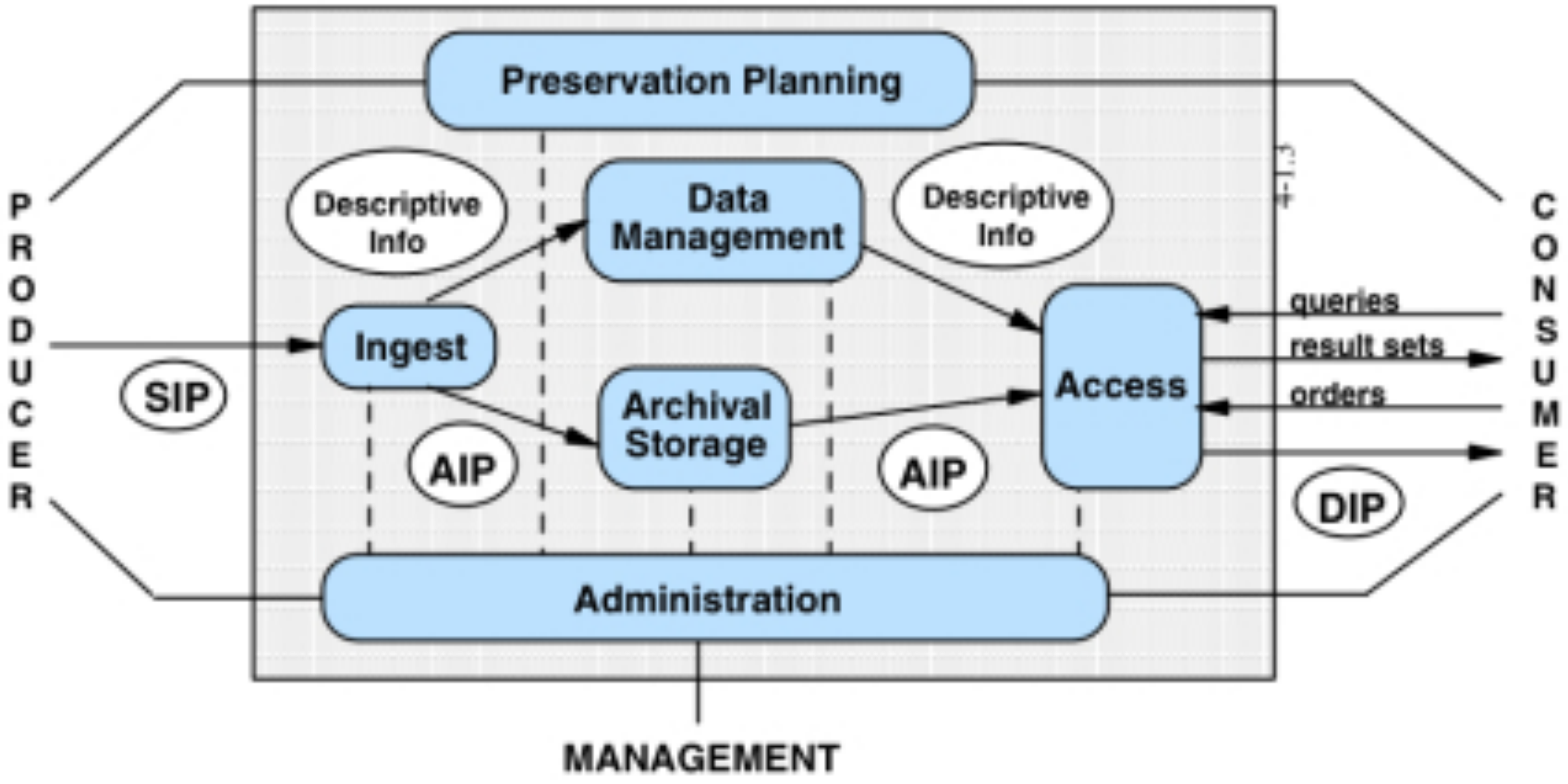
Infrastructures & Architectures for Data Curation

LIS 598
Data Curation 2

Agenda

- Conceptual model: OAIS
- Architectures, stacks, and layers
- Dataverse
    - Software and hardware configurations
    - Use case for Dataverse at QDR
    - Alternative architecture at data.Gov for geospatial data


OAIS
Preservation Planning
Data Management
Ingest
Archival Storage
Access
Administration
PRODUCER
CONSUMER
MANAGEMENT


Preservation Planning
Descriptive Info
Data Management
Descriptive Info
P R O D U C E R
Ingest
SIP
AIP
Archival Storage
AIP
Access
queries
result sets
orders
C O N S U M E R
DIP
Administration
MANAGEMENT

- System Architecture: The configured software, hardware, and protocols that constitute an information system

- “stack” - this is the notion that there are a set of compatible technologies that enable complexity.

    - LAMP stack - Linux (operating system), Apache (servers), MySQL (databases), Python (programming language for operations)

- Layers: The technologies that combine to achieve a certain class of system performance (data layer, web layer, application layer)

- **Data repositories exist at the application layer**

- Data repositories are differentiated by their underlying architecture. Many architectures are optimized for one kind of data vs another.

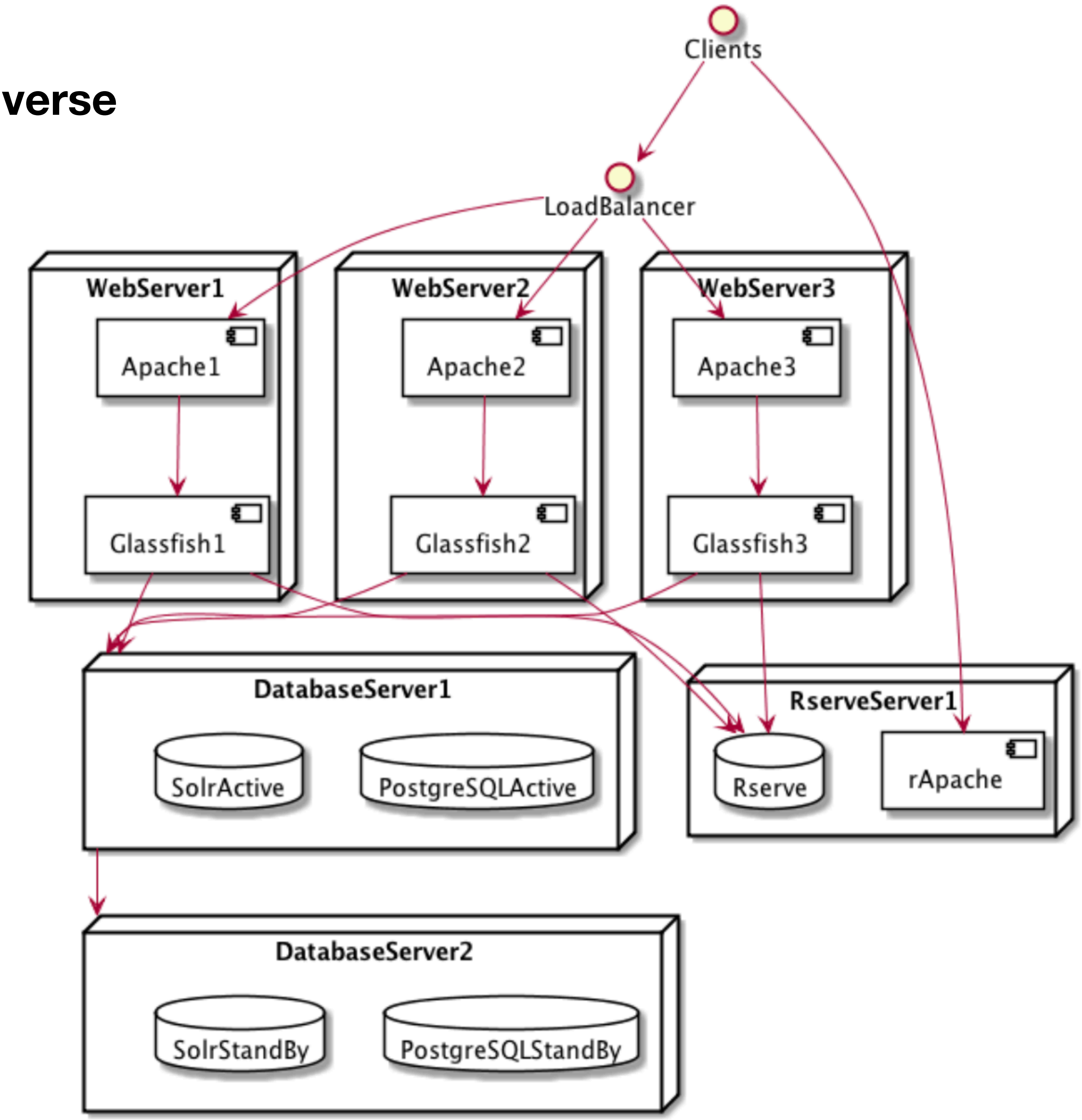
The Dataverse® Project

An open-source community project for storing and sharing social science data

- Dataverse software architecture:
  - **Linux**: RHEL/CentOS is highly recommended since all development and QA happens on this distribution.
  - **Glassfish**: a Java EE application server to which the Dataverse application (war file) is to be deployed.
  - **PostgreSQL**: a relational database.
  - **Solr**: a search engine. A Dataverse-specific schema is provided.
  - **SMTP server**: for sending mail for password resets and other notifications.
  - **Persistent identifier service**: DOI and Handle support are provided. Production use requires a registered DOI or Handle.net authority.
  - Other related software: Authentication systems; Alternative web servers (Apache) for HTTP traffic and load balancers…

- Dataverse hardware requirements:
  - Minimum: two 2.8 GHz processors, 8 GB of RAM and 50 GB of disk. (most of our laptops can run that!)
  - Most builds: 128gb of RAM across multiple machines; multiple CPUs at 2.8+; 1TB of disk for staged data, and multiple AWS S3 buckets of 1TB or more for offline data storage.

Harvard's Dataverse



- System architectures can be realized in many different configurations

- Configuration simply means how we set up different components (servers, databases, and webpages) to be connected to one another.

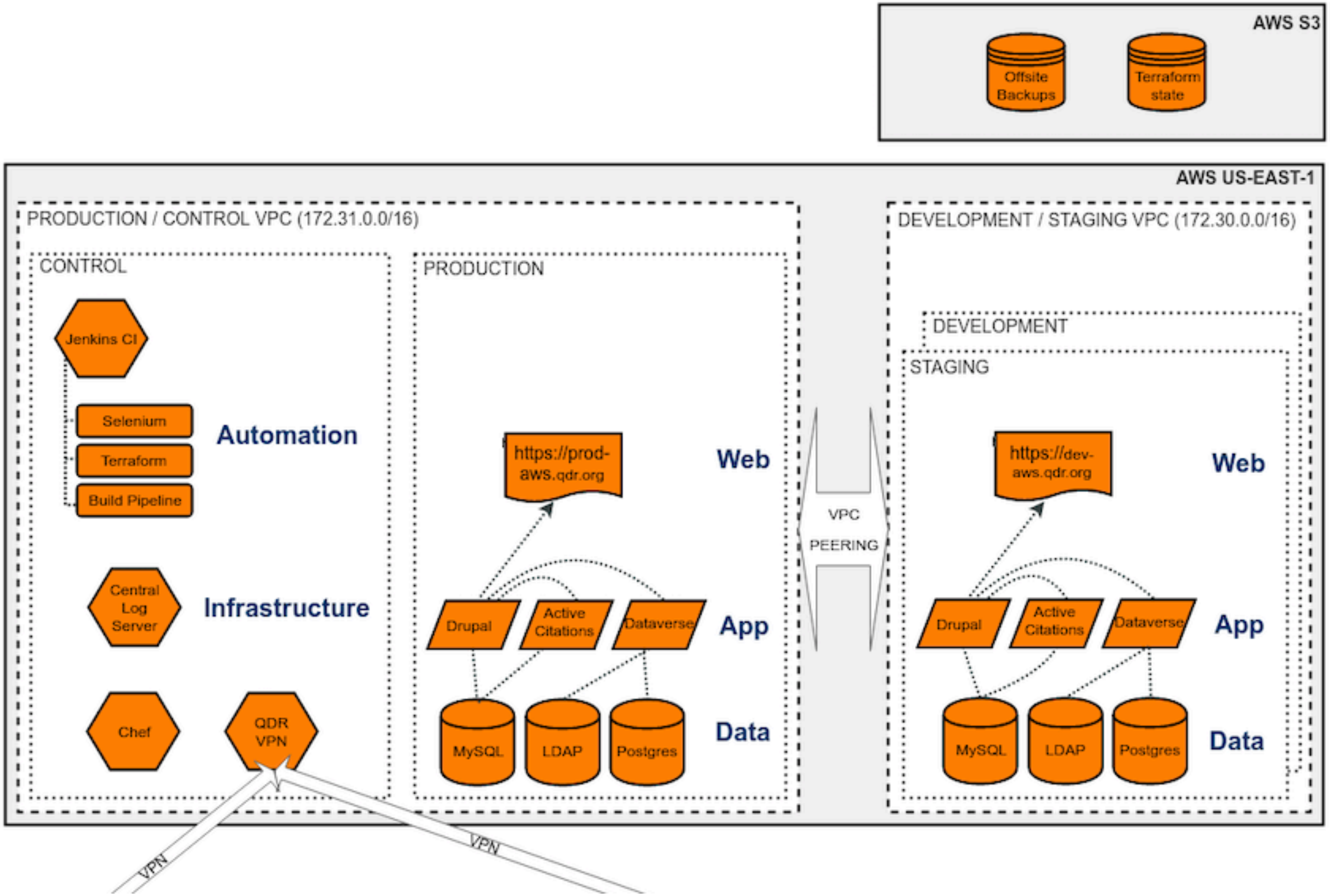
- Configuration of a data repository's architecture can be based on

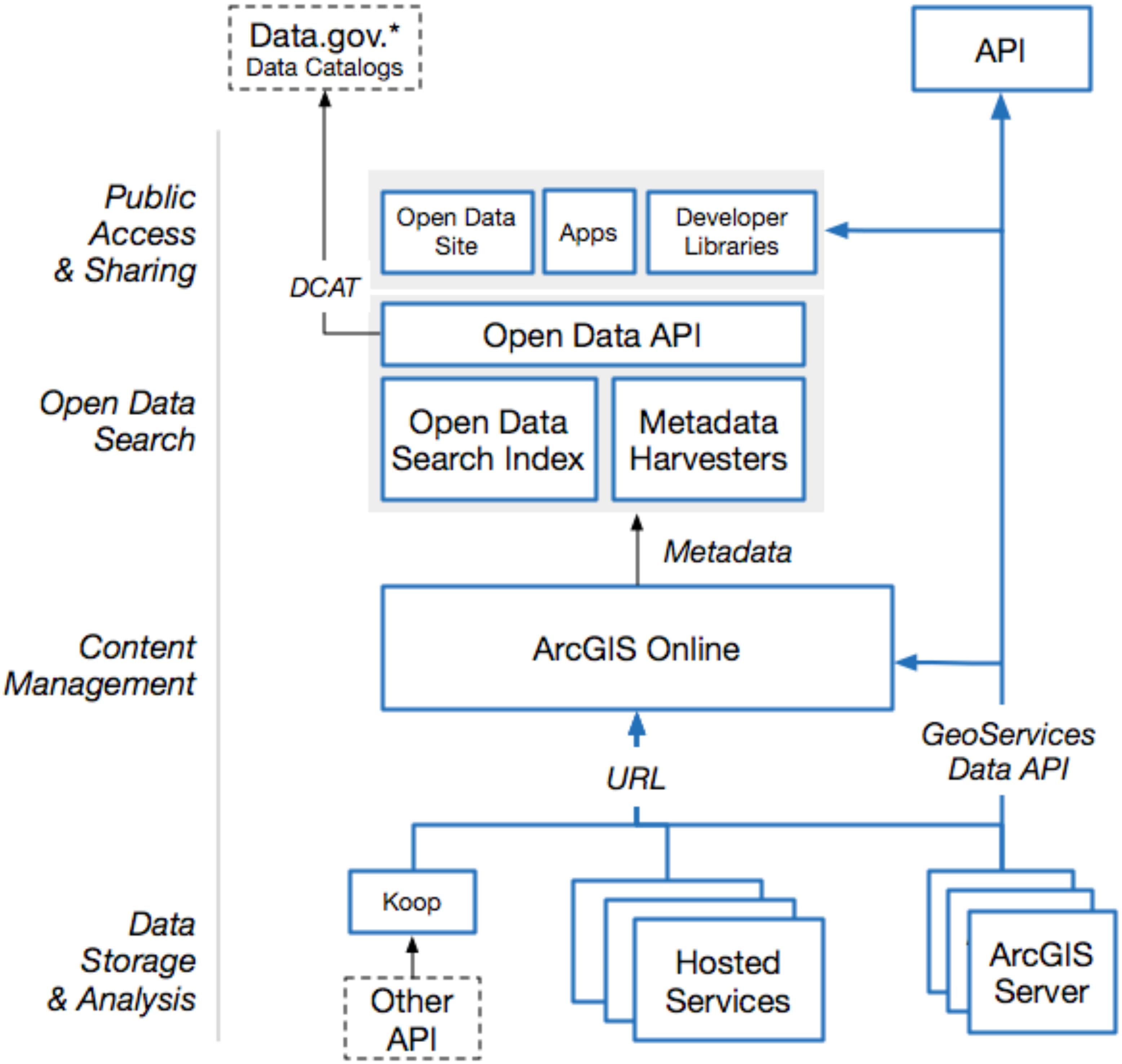
    - Controls of data access

    - Security

    - Performance

    - Redundancy


AWS S3
Offsite Backups
Terraform state
AWS US-EAST-1
PRODUCTION / CONTROL VPC (172.31.0.0/16)
CONTROL
Jenkins CI
Selenium
Terraform
Build Pipeline
Automation
Central Log Server
Infrastructure
Chef
QDR VPN
PRODUCTION
https://prod-aws.qdr.org
Web
Drupal
Active Citations
Dataverse
App
MySQL
LDAP
Postgres
Data
VPC PEERING
DEVELOPMENT / STAGING VPC (172.30.0.0/16)
DEVELOPMENT
STAGING
https://dev-aws.qdr.org
Web
Drupal
Active Citations
Dataverse
App
MySQL
LDAP
Postgres
Data
VPN
VPN



Data.gov.*
Data Catalogs

API

Public
Access
& Sharing

Open Data
Site

Apps

Developer
Libraries

DCAT

Open Data API

Open Data
Search

Open Data
Search Index

Metadata
Harvesters

Metadata

Content
Management

ArcGIS Online

GeoServices
Data API

URL

Data
Storage
& Analysis

Koop

Other
API

Hosted
Services

ArcGIS
Server

- With architecture diagrams we are attempting to map the different components of an information system

- For data repositories that exist at the application layer - there is a need to understand exactly how each of these different technologies interact so that we can advocate for data being securely managed and preserved