

Data Center Basics

WHAT A DATA CENTER IS · CORE ELEMENTS · FOUR TYPES · DESIGN REQUIREMENTS · AVAILABILITY & RESILIENCE

STUDY & REVIEW

01 MUST KNOW

if you remember five things, remember these

- 01 A data center is a facility, not a service: a physical facility** that houses servers, storage, network, and communication equipment and **centralizes** an organization's IT operations to store, process, and distribute data. Its whole job is to keep one of an organization's most critical assets — its **data** — available.
- 02 Four core elements: core components** (servers, racks, network/firewall, storage), the **facility** (temperature-controlled space), the **support infrastructure** (power, cooling, physical security), and the **operations staff** who run and maintain it.
- 03 Four types: enterprise** (built and run in-house), **colocation** (rent space/racks — you manage your servers, the provider manages the building), **cloud-based** (pay-as-you-go, the CSP owns the infrastructure), and **managed services** (outsourced operation). Data centers are classified by **technology** and **ownership**.
- 04 Two performance measures: availability** = $\text{uptime} \div (\text{uptime} + \text{downtime})$ — how much of the time the facility is operational; **resilience** = how fast it recovers, set by **RTO** (time to restore) and **RPO** (acceptable data loss). **Power and cooling** have the biggest impact on availability.
- 05 Redundancy is targeted, not total:** duplicate only **high-risk** equipment; levels run **N+1** (one shared spare) → **2N** (a full second set) → **2N+1** (full second set plus one shared spare). Pair redundancy with **preventive, predictive, and scheduled** maintenance.

02 KEY TERMS

TERM	DEFINITION
Data center	A physical facility that houses servers, storage, network, and communication equipment and centralizes an organization's IT operations for storing, processing, and distributing data
Core components	The hardware- and software-related equipment that stores the data — servers, routers, racks, network and firewall infrastructure, and storage — kept under highly restricted access
Support infrastructure	Systems that keep the facility running: redundant power (generators, UPSs), physical security (CCTV, access controls), and environmental controls (AHUs, cooling)
Operations staff	Trained personnel who oversee and monitor operations, maintain the infrastructure and facility, and troubleshoot faults to keep downtime near zero
Colocation data center	A facility that rents physical space or racks to an organization, which manages its own servers while the provider maintains the building and utilities
Fault tolerance	The ability of a system to keep working uninterrupted, without breaking down, even when a major component fails; applies to equipment and to people
Redundancy	A design in which systems or components are duplicated so a backup can take over when one fails
UPS	Uninterruptible power supply — battery power sized to carry the load for about 5–20 minutes during an outage while continuously filtering and conditioning power
PDU	Power distribution unit — equipment in the electrical backbone that distributes power and integrates circuit breakers
HVAC	Heating, ventilation, and air-conditioning — the localized system that controls temperature, humidity, and airflow and removes air pollutants
Availability	The share of time a data center is operational (not in failure mode or maintenance), calculated as $\text{uptime} \div (\text{uptime} + \text{downtime})$

TERM	DEFINITION
Resilience	A data center's ability to recover quickly and continue operating after a failure, measured by RTO and RPO

03 THE FOUR TYPES OF DATA CENTERS

classified by technology and ownership

TYPE	WHAT IT IS	WHO RUNS THE FACILITY
Enterprise	Built and owned by one organization for its own applications; the org selects the site, customizes the design, and runs daily operations	The organization itself
Colocation	The organization rents physical space or a few racks and manages its own servers and storage	Provider maintains the building and utilities
Cloud-based	Pay-as-you-go infrastructure reachable from anywhere; the customer still owns app security, OS, firewalls, accounts, and network	The cloud service provider (CSP)
Managed services	Outsourced data center — the org leases access and hands operations to a third party; dedicated infrastructure for its apps	A third-party vendor

04 CRITICAL SUPPORT INFRASTRUCTURE

five systems that keep a facility alive

SYSTEM	WHAT IT DOES
Power source	Distributes constant, uninterrupted, surge-safe power — UPSs, diesel generators, PDUs, and integrated data center units (IDUs)
HVAC	Localized air-cooling that controls temperature, humidity, and airflow and removes air pollutants
Fire protection	Alarm and notification systems that detect fire early and help prevent it from spreading
Security	Systems that monitor, authorize, and restrict access to the data center
Raised floor	Provides space for cold airflow beneath the floor and secures power and data cables

05 THE REDUNDANCY SPECTRUM

more redundancy = higher availability and cost

01 N — just enough for full IT load, zero spare → **02 N+1** — one shared backup (11 for a 10-unit load) → **03 2N** — a full second set (20 units) → **04 2N+1** — full second set plus one shared spare (21 units)

06 AVAILABILITY VS. RESILIENCE

the two performance measures every operator tracks

AVAILABILITY <i>is it up right now?</i> — Share of time the facility is operational — = uptime ÷ (uptime + downtime) — Raised by redundancy (N+1, 2N, 2N+1) — Power and cooling matter most	RESILIENCE <i>how fast does it come back?</i> — Ability to recover and continue after a failure — RTO — time allowed to restore service — RPO — acceptable amount of data lost — Lower RTO/RPO ⇒ more redundant equipment
---	---

- ✗ “Cloud-based means the provider handles all your security.” — In a cloud-based data center the **customer** still owns application security, the operating system, firewalls, accounts, and the network; the CSP manages the physical infrastructure.
- ✗ “In colocation, the provider manages your servers.” — The **customer** manages its own servers and storage; the colocation provider only maintains the building infrastructure and utilities.
- ✗ “Make every piece of equipment redundant.” — Plan redundancy only for **high-risk** items; duplicating low- and medium-risk gear just raises cost and overloads operations.
- ✗ “2N and 2N+1 are the same.” — **2N** is a full duplicate set; **2N+1** adds one more shared spare that can back up any pair.
- ✗ “A UPS carries the data center through a long outage.” — A UPS is sized for only about **5–20 minutes** — long enough to condition power and bridge to the **generators**, which cover extended outages.
- ✗ “Availability and resilience measure the same thing.” — **Availability** measures how much time the facility is up (uptime vs. downtime); **resilience** measures how fast it recovers (**RTO**) and how much data it could lose (**RPO**).
- ✗ “Fault tolerance is only about hardware and software.” — It also applies to **people** (operations staff), and it means the system keeps running even during a **major component failure**, not just minor glitches.
- ✗ “Lower RTO is always better, whatever it costs.” — A lower RTO demands more redundant equipment; redundancy must **balance** setup and maintenance cost against the RTO target — you don't duplicate everything.

08 SELF-QUIZ

answers are upside-down below

- Q1** Name the four core elements of a data center.

Q2 Which data center type rents out physical space or a few racks while the tenant manages its own servers?

Q3 In a cloud-based data center, who remains responsible for application security, the operating system, firewalls, accounts, and the network?

Q4 Write the calculation that measures a data center's availability.

Q5 A facility needs 10 diesel generators at full load. How many are installed to reach 2N redundancy?
- Q6** Name the three types of maintenance a data center relies on.

Q7 Which resilience parameter defines the maximum acceptable amount of data that could be lost after a failure?

Q8 Which two primary infrastructure elements have the biggest impact on a data center's availability?

Q9 For which risk category should you always plan redundancy?

Q10 For roughly how long should a UPS be sized to carry the load during a power outage?

ANSWER KEY (rotate the page)

Q1 core components, facility, support infrastructure, and operations staff **Q2** colocation **Q3** the customer **Q4** uptime ÷ (total uptime + total downtime) **Q5** 20 **Q6** preventive, predictive, and scheduled **Q7** RPO (recovery point objective) **Q8** power and cooling **Q9** high-risk items **Q10** about 5 to 20 minutes