
aws

Why data pipelines on cloud are cheaper than on-prem solution?

Data pipelines on the cloud are often cheaper than on-premises solutions due to several reasons:

- **Infrastructure Costs**: Cloud platforms eliminate the need to invest in physical hardware, reducing both upfront capital expenses and ongoing maintenance costs.
- **Scalability**: Cloud services allow for on-demand scalability. Instead of over-provisioning resources, businesses can scale up or down based on actual usage, optimizing costs.
- **Operational Efficiency**: Cloud providers handle much of the routine maintenance, patching, and updates, reducing the manpower required for these tasks on-premises.
- **Resource Utilization**: Cloud platforms typically offer better resource utilization and multi-tenancy benefits, allowing users to only pay for the resources they consume.
- **Reduced Overhead**: On-premises solutions often require dedicated IT teams for setup, maintenance, and troubleshooting. Cloud solutions can reduce or eliminate these overhead costs.
- **Flexibility**: The cloud offers various pricing models, such as pay-as-you-go, reserved instances, or spot instances, enabling cost optimizations based on workload patterns.
- **Innovation Pace**: Cloud providers continuously introduce new features, tools, and integrations, often at no additional cost. This can lead to more efficient and cost-effective pipeline designs over time.
- **Disaster Recovery & Redundancy**: Implementing redundancy and disaster recovery on-premises can be expensive. Cloud providers often offer these features at a fraction of the cost.

Common Services by AWS - Azure - GCP

Grow Data Skills

	Azure	AWS	Google
Available Regions	Azure Regions	AWS Regions and Zones	Google Compute Regions & Zones
Compute Services	Virtual Machines	Elastic Compute Cloud (EC2)	Compute Engine
App Hosting	Azure Cloud Services	Amazon Elastic Beanstalk	Google App Engine
Serverless Computing	Azure Functions	AWS Lambda	Google Cloud Functions
Container Support	Azure Container Service	EC2 Container Service	Container Engine
Scaling Options	Azure Autoscale	Auto Scaling	Autoscaler
Object Storage	Azure Blob Storage	Amazon Simple Storage (S3)	Cloud Storage
Block Storage	Azure Managed Storage	Amazon Elastic Block Storage	Persistent Disk
Content Delivery Network (CDN)	Azure CDN	Amazon CloudFront	Cloud CDN
SQL Database Options	Azure SQL Database	Amazon RDS	Cloud SQL
NoSQL Database Options	Azure DocumentDB	AWS DynamoDB	Cloud Datastore
Virtual Network	Azure Virtual Network	Amazon VPC	Cloud Virtual Network
Private Connectivity	Azure Express Route	AWS Direct Connect	Cloud Interconnect
DNS Services	Azure Traffic Manager	Amazon Route 53	Cloud DNS
Log Monitoring	Azure Operational Insights	Amazon CloudTrail	Cloud Logging
Performance Monitoring	Azure Application Insights	Amazon CloudWatch	Stackdriver Monitoring
Administration and Security	Azure Active Directory	AWS Identity and Access Management (IAM)	Cloud Identity and Access Management (IAM)
Compliance	Azure Trust Center	AWS CloudHSM	Google Cloud Platform Security
Analytics	Azure Stream Analytics	Amazon Kinesis	Cloud Dataflow
Automation	Azure Automation	AWS Opsworks	Compute Engine Management
Management Services & Options	Azure Resource Manager	Amazon Cloudformation	Cloud Deployment Manager
Notifications	Azure Notifcation Hub	Amazon Simple Notification Service (SNS)	None
Load Balancing	Load Balancing for Azure	Elastic Load Balancing	Cloud Load Balancing

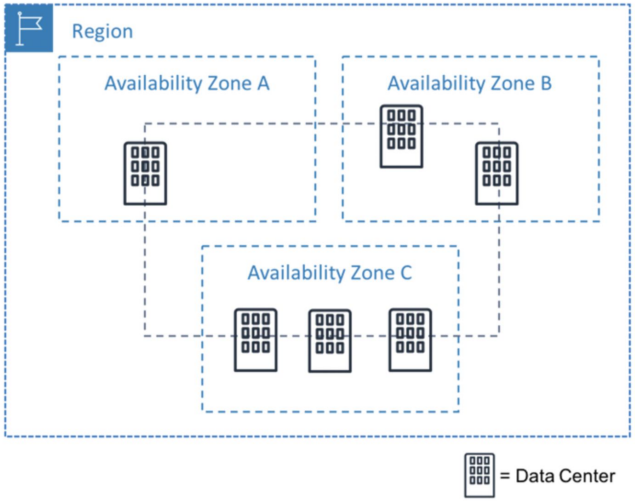
AWS Region, Availability Zone and Data Centers

Grow Data Skills

- **Region**: An AWS Region is a geographic area that contains two or more Availability Zones. AWS Regions are completely independent of one another and are designed to be completely isolated from each other, to ensure the highest levels of data privacy and resilience. A region represents a large area, such as the US West (Oregon) or Europe (Ireland).

- **Availability Zone:** An Availability Zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity in an AWS Region. Each AZ is designed to be isolated from failures in other AZs and to provide inexpensive, low-latency network connectivity to other AZs in the same region.

- **Data Center:** A physical facility that houses the servers and networking equipment necessary to provide the various AWS services. Multiple data centers make up an Availability Zone, and multiple Availability Zones make up a Region.



AWS S3

Grow Data Skills

Amazon S3 (Simple Storage Service) is one of the foundational services in the AWS suite and is widely used by businesses and individuals to store and retrieve any amount of data, at any time, from anywhere.

Overview of AWS S3:

- **Object Storage:** S3 is an object storage service, meaning it is designed to store unstructured data (like photos, videos, backups, etc.) as objects within resources called "buckets".
- **Durability and Availability**: AWS S3 is designed for 99.999999999% (11 9's) durability over a given year. This ensures that your data remains safe and intact.
- **Scalability**: There's no limit to the amount of data you can store in S3, and it's designed to handle high request rates and traffic.
- **Data Organization**: Data in S3 is organized into buckets (similar to directories) and objects (files).
- **Versioning**: S3 supports versioning, allowing you to retain, retrieve, and restore every version of every object in your bucket.
- **Security**: Offers features like bucket policies, ACLs (Access Control Lists), and server-side encryption (SSE) for data. Integrated with AWS Identity and Access Management (IAM) for access control.
- **Event Configuration**: You can set up event notifications to trigger workflows, alerts, or other automated processes based on changes to your data.

AWS S3 Pricing Factors

Grow Data Skills

AWS S3 pricing is based on several factors:

- **Storage**: You're billed per GB per month based on the amount of data stored.
- **Requests**: Costs associated with the number and type of requests made (GET, PUT, COPY, etc.).
- **Data Transfer**: While transferring data into S3 is typically free, transferring data out of S3 to the internet or other AWS regions incurs charges.
- **Additional Features**: Features like versioning, monitoring with CloudWatch, data transfer acceleration, and others might have associated costs.
- **Storage Management**: Using features like S3 Inventory, S3 Analytics, and S3 Object Tagging will also influence the total cost.

S3 command with AWS CLI

Grow Data Skills

The AWS Command Line Interface (CLI) is a powerful tool that allows users to interact with AWS services, including S3, directly from the command line. Here's a list of some commonly used AWS S3 CLI commands:

- **Configuration:**
  - **aws configure**: Setup the CLI with your AWS credentials, default region, and desired output format.

- **Bucket Operations:**
  - **aws s3 ls:** List all buckets.
  - **aws s3 mb s3://my-bucket-name:** Create a new bucket.
  - **aws s3 rb s3://my-bucket-name:** Delete a bucket.

- **File and Folder Operations:**
  - **aws s3 ls s3://my-bucket-name:** List contents of a bucket.
  - **aws s3 cp localfile.txt s3://my-bucket-name/:** Copy a local file to a bucket.
  - **aws s3 cp s3://my-bucket-name/file.txt localfile.txt:** Copy a file from a bucket to the local system.
  - **aws s3 mv localfile.txt s3://my-bucket-name/:** Move a local file to a bucket (removes the local file after copying).
  - **aws s3 rm s3://my-bucket-name/file.txt:** Delete a file from a bucket.