All about Data Engineering

AWS S3 - Simple Storage Service

by Sachin Chandrashekhar

Data Engineering Hub

https://masterclass.sachin.cloud →

What is AWS S3?

Amazon Simple Storage Service (S3) is an object storage service that provides highly scalable, secure, and durable storage for data of all sizes. It is used for everything from backups to big data processing.

Data Engineering Hub

- Sachin Chandrashekhar

How S3 Stores Data

S3 stores data as objects in buckets. Each object is made up of data, metadata, and a unique identifier (key). This flat, key-based architecture makes S3 a flexible solution for data storage.

Data Engineering Hub

- Sachin Chandrashekhar

Global Availability of S3

S3 operates across multiple AWS regions and Availability Zones (AZs), ensuring that your data can be globally available with replication and redundancy to meet compliance and disaster recovery needs.

https://masterclass.sachin.cloud

S3 Use Cases

S3 is used for a wide range of use cases including data lakes, backups, content distribution, archival storage, and web hosting. Its flexibility makes it a key component in data engineering pipelines.

Data Engineering Hub

- Sachin Chandrashekhar

How S3 Secures Data

S3 secures data using encryption at rest and in transit. Features like access control lists (ACLs), bucket policies, and integration with AWS IAM ensure granular control over who can access your data.

https://masterclass.sachin.cloud

Data Engineering Hub

- Sachin Chandrashekhar

S3 Cost Management

Cost management in S3 involves choosing the right storage class, leveraging lifecycle policies, using data compression, and regularly analyzing access patterns to optimize costs without compromising performance.

https://masterclass.sachin.cloud →

Data Engineering Hub

- Sachin Chandrashekhar

S3 Data Lakes

AWS S3 serves as a scalable foundation for building secure data lakes. It integrates with AWS Lake Formation and Amazon Athena to enable efficient data ingestion, storage, and querying, making it ideal for large-scale data engineering.

https://masterclass.sachin.cloud →

Data Engineering Hub

- Sachin Chandrashekhar

S3 Intelligent-Tiering

Intelligent-Tiering optimizes storage costs by automatically moving data between frequent and infrequent access tiers based on usage patterns, allowing data engineers to store large volumes of data without overpaying for storage.

https://masterclass.sachin.cloud →

S3 and AWS Lambda Integration

S3 can trigger AWS Lambda functions for events such as object uploads or deletions. Data engineers can automate workflows by processing data in real-time, optimizing processes like ETL, file validation, and data ingestion.

S3 and AWS Glue Integration

AWS Glue can crawl and transform data stored in S3, making it easier for data engineers to prepare and catalog data for analytics. Glue's serverless capabilities complement S3's scalability for large-scale data processing.

S3 and Amazon Athena

Athena allows data engineers to run SQL queries directly on data stored in S3 without needing a database. This serverless approach to querying is highly scalable and cost-effective for big data analytics.

Data Engineering Hub

- Sachin Chandrashekhar

S3 and Amazon EMR Integration

Amazon EMR processes vast amounts of data stored in S3 using big data frameworks like Apache Spark and Hadoop. This integration enables scalable, cost-effective data processing pipelines for data engineering use cases.

https://masterclass.sachin.cloud →

S3 Cross-Region Replication (CRR)

Cross-Region Replication automatically replicates objects across different AWS regions, ensuring data redundancy and compliance with data locality laws. Itâ€™s a powerful tool for global data engineering teams.

Data Engineering Hub

- Sachin Chandrashekhar

S3 Event Notifications

S3 Event Notifications can trigger Lambda functions, send messages to SQS, or initiate SNS notifications based on object-level events. This allows data engineers to build real-time data processing workflows.

https://masterclass.sachin.cloud →

S3 and Data Cataloging with Glue

Glue's Data Catalog integrates seamlessly with S3, enabling data engineers to catalog, search, and discover datasets stored in S3, simplifying data management for analytics and reporting.

Data Engineering Hub

- Sachin Chandrashekhar

S3 Versioning for Data Recovery

Versioning in S3 allows data engineers to keep multiple versions of objects, preventing accidental deletion or modification. This feature is critical for maintaining data integrity in engineering workflows.

S3 and Amazon Redshift Spectrum

Redshift Spectrum allows data engineers to query data in S3 directly from Amazon Redshift, combining the scalability of S3 with the powerful analytics capabilities of Redshift for large-scale data analysis.

Data Engineering Hub

- Sachin Chandrashekhar

S3 Multipart Upload for Large Data Sets

Multipart Upload in S3 enables efficient uploading of large datasets by splitting them into smaller parts. This ensures faster, more reliable uploads, ideal for data engineers working with large files.

https://masterclass.sachin.cloud

Data Engineering Hub

- Sachin Chandrashekhar

S3 Encryption Options

S3 offers server-side and client-side encryption to protect data at rest. Data engineers can choose from options like SSE-S3, SSE-KMS, or SSE-C to meet specific security requirements.

https://masterclass.sachin.cloud →

Data Engineering Hub

- Sachin Chandrashekhar

S3 Glacier for Archival Storage

S3 Glacier provides cost-effective storage for data that is accessed infrequently, making it a go-to solution for long-term data archival in data engineering.

Data Engineering Hub

- Sachin Chandrashekhar

S3 and AWS CloudTrail for Data Governance

CloudTrail records API calls made to S3, allowing data engineers to track and audit activities for security and compliance, ensuring full governance over their data assets.

https://masterclass.sachin.cloud

Data Engineering Hub

- Sachin Chandrashekhar

S3 Strong Read-After-Write Consistency

S3 offers strong read-after-write consistency for all operations, eliminating the need for additional code or architecture to ensure data consistency across distributed engineering applications.

https://masterclass.sachin.cloud

All about Data Engineering

Find this useful? like and share this post with your friends.

by Sachin Chandrashekhar
https://masterclass.sachin.cloud

Save