

# Armazenamentos de Arquivos

- ◊ Não se trata apenas de um repositório de documentos!
- ◊ Dados “vivos” são mantidos em sistemas de arquivos
- ◊ Apresentam características como:
  - ◊ Versionamento
  - ◊ Segurança: controle de acesso, criptografia etc.
  - ◊ Ciclo de Vida
- ◊ Características Relacionadas a “Big Data”:
  - ◊ Particionamento
  - ◊ Escalabilidade
- ◊ Exemplos
  - ◊ S3
  - ◊ HDFS
  - ◊ Azure Blob Storage

# Visão Geral do AWS S3

- ◇ O Amazon S3 permite que as pessoas armazenem objetos (arquivos) em buckets
- ◇ Buckets devem ter um nome único global
- ◇ Objetos (arquivos) tem uma chave. A chave é o caminho completo:
  - ◇ <bucket>/vendas.csv
  - ◇ <bucket>/pasta1/pasta2/vendas.csv
- ◇ Isso é útil e interessante quando olharmos partições
- ◇ Tamanho máximo de um objeto é de 5TB
- ◇ Tags de Objetos (chave / valor, até 10), úteis para segurança e ciclo de vida



# AWS S3 para Ciência de Dados

- ◊ Backbone para muitos serviços de ML do AWS (ex: SageMaker)
- ◊ Criar um Data Lake
  - ◊ Tamanho infinito, sem provisionamento
  - ◊ Durabilidade 99,999999999%
  - ◊ Armazenamento (S3) desacoplado do processamento (EC2, Amazon Athena, Amazon Redshift Spectrum, Amazon Rekognition e AWS Glue)
- ◊ Arquitetura centralizada
- ◊ Armazenamento de objetos: suporta qualquer tipo de arquivo
- ◊ Formatos comuns para Eng. Dados: CSV, JSON, Parquet, ORC, Avro, Protobuf



# AWS S3: Particionamento



- ◆ Padrões para acelerar consultas em intervalos (ex: AWS Athena)
- ◆ Por data: `s3://<bucket>/vendas/ano/mês/dia/hora/venda_00.csv`
- ◆ Por produto: `s3://<bucket>/vendas/234565/venda_00.csv`
- ◆ Você pode definir qual estratégia de particionamento você quer
- ◆ O particionamento de dados pode ser feito pelas próprias ferramentas do AWS (Glue)