

***Armazenamento e Processamento Massivo e Distribuído de Dados***

Introdução aos sistemas distribuídos. Princípios de processamento e de volumes de dados massivos. Sistemas de arquivos distribuídos. Modelo MapReduce. Balanceamento de carga. Replicação e redundância de dados e metadados. Gestão de metadados. Sincronização e disponibilidade. Operações de leitura, escrita e listagem de arquivos distribuídos. Soluções em Processamento Distribuído. Técnicas de compressão, otimização e particionamento de dados.

***Arquiteturas e Serviços de Data Lakes e Data Warehousing***

Conceitos de Data Lake e Data Warehouse. Dados Estruturados x Dados Não Estruturados. Logical Data Lake. Data ponds. Blocos funcionais em uma arquitetura de dados. Schema on-write x Schema on-read. Sincronização de fluxos independentes e dependentes. Arquitetura Corporativa - Enterprise Data Hub. Construção de arquiteturas com alta disponibilidade (HA). Construção de dimensões em DW. Engines de Data Warehousing. Orquestração de transformações em DW. Organização de camadas em um Data Lake. Data Lake On-premise x Data Lake na Nuvem.

***Bancos de Dados NoSQL***

Motivações para bancos NoSQL. Conceitos e características dos bancos de dados NoSQL Principais benefícios de um banco de dados NoSQL. Principais tipos de Bancos NoSQL. Key-value databases. Document databases. Column Family Databases. Graph Databases. Principais diferenças arquiteturais entre os tipos. Principais casos de uso de cada tipo.

***Bancos de Dados Relacionais e Linguagem SQL***

Fundamentos de bancos de dados. Modelagem de bancos de dados relacionais (Diagrama ER). Linguagem SQL: Data Manipulation Language (DML) e Data Definition Language (DDL).

***Camadas e Serviços de Consumo de Dados***

Produtores e consumidores de dados. Dados de domínio como produto. Camadas de dados. Critérios para determinar a camada de dados. Barramentos de mensageria de dados. Particionamento e técnicas de cache para acessos de alta frequência. Planos de execução. Consumo de alta volumetria de dados. Camadas de consumo por serviços. Autenticação e controle de acessos. Design convergente para dados e plataforma self-service.

***Computação em Nuvem***

Aspectos da Computação em Nuvem: conceitos, tipos, utilização, fornecedores, utilização corporativa. Soluções corporativas de SaaS, PaaS e IaaS. Aspectos de segurança da informação. Principais plataformas de Colaboração e Comunicação. Arquiteturas de Cloud: MultiCloud e Cloud híbridas.

***Cultura e Práticas Devops***

A cultura DevOps. Integração contínua e entrega contínua. Estratégias de deploy. Projeto de pipeline para build e deployment. Automação de testes. Infrastructure as Code (IaC). Ferramentas e infraestrutura do ambiente integrado DevOps: Containers, Docker, Kubernetes e OpenShift.

***Governança de Dados***

Contexto organizacional de dados. Conceitos de Governança de Dados (GD). Framework DMBOK. Políticas, padrões e procedimentos aplicados aos dados: Data Stewardship, Data Owners, Dados Mestres, Dados Referência, Metadados, Data Catalog. Processo de implantação de GD. Modelos de maturidade de dados. GD aplicada em leis de Proteção (LGPD-GDPR). Compliance e Risk Assessment. GD 2.0: Ética nos dados, Agilidade em GD, Gerência de Mudanças.

***Ingestão e Catalogação de Dados***

Conceitos e técnicas de ingestão de dados. Ingestão Change Data Capture (CDC). Slowly Change Dimension. Técnicas de ingestão baseadas em diferenças temporais. Conectores de migração em tempo-real. Ingestão por ETL / ELT clássico. Fluxo de dados real Time e CEP (Complex Event Processing) Tratamento de eventos. Mecanismos de inferência de tipagem. Técnicas de consistência e harmonização de tipagem. Schema assert. Schema-registry. Monitoramento de camadas de ingestão de dados. Organização de áreas de dados RAW. Ferramentas e tecnologias para ingestão de dados. Setup de infraestrutura de rede.

***Otimização, Monitoramento e Operação em Bancos DE***

Indexação. Análise de desempenho. Otimização e tuning (relacional e NoSQL). Benchmarks para SGBD. Construção de soluções de monitoramento de arquiteturas de dados. Indicadores, alarmes e métricas de acompanhamento. Serviços de alertas baseados em cloud. Recuperação de falhas. Sistemas de backup. Bancos de dados ativos (triggers). Ferramentas de visualização para monitoramento.

***Preparação, Orquestração e Fluxos de Dados***

Montagem do conjunto de dados. Tipos de dados. Melhoramento e enriquecimento. Preparação: eliminação de dados irrelevantes, granulação e agregação, consistência, poluição, análise de domínios de atributos, integridade, concordância, duplicação e redundância. Feature Engineering. Combinando dados de múltiplas fontes. ELT x ETL. Transformação e transferência de dados. Ferramentas para preparação de dados. Ferramentas de orquestração. Escalonamento de jobs baseados em eventos. Reprocessamento em serviços de migração. Resiliência em migração. Orquestradores open-source. Orquestradores baseados em cloud. Implementação de fluxos de correção de dados em expurgo.

***Processamento de Fluxos Discretos e Contínuos de Dados***

Processamento em lotes. Processamento de dados gerados por eventos. Spark, spark-streaming, Kafka, serviços gerenciados em cloud, soluções serverless e orientadas a eventos. Catalogação de streaming para controle de jobs. Configuração do Spark.

***Provisionamento de Infraestrutura e Pipelines***

Fundamentos da Arquitetura de dados escaláveis. Infraestrutura de soluções de dados como código (IAC). Arquitetura Escalável: Dimensionamento, monitoramento, segurança, cálculo de capacidade, disponibilidade e desempenho. Escalonamento vertical e horizontal. Projeto de Arquitetura de dados escalável. Criação de pipelines automáticos. Deploy automático de componentes de arquitetura de dados. Uso de Docker e SDKs para gestão de pipeline. Gestão e ciclo de vida de pipeline de Engenharia de Dados. Introdução a DataOps.

***Python para Engenharia de Dados***

Fundamentos da Linguagem Python. Bibliotecas para manipulação de dados. Bibliotecas para visualização de dados. Conexões com Bancos de Dados e arquivos. Construção de módulos de migração. Operações de processamento e extração. Manipulação de dados com APIs. Implementação de módulos serverless. Processamento multi-threading e distribuídos.

***Segurança em Armazenamento, Tratamento e Consumo de Dados***

Criptografia e gestão de chaves criptográficas. Serviços e técnicas de autenticação. Tratamento de dados sensíveis e mascaramento. Planejamento e regras de rede. Gestão de acessos de usuários e serviços. Implementação de padrões e processos compatíveis com LGPD / GDPR. Análise de riscos. A necessidade da adoção dos controles internos. Conceitos de auditoria.