Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 1. Computação em Nuvem

Aula 1.1. O que é computação em nuvem?

Prof. Taynan Ferreira

Nesta aula

- Definição de computação em nuvem.
- Principais vantagens da computação em nuvem.

Definição

- “A disponibilização sob demanda de poder computacional, banco de dados, armazenamento, aplicações e outros recursos de TI através de uma plataforma de serviços via internet com cobrança conforme o uso”

Overview of Amazon Web Services

AWS Whitepaper

Definição

- On-Premise x Computação em Nuvem.
- On-Premise x Cloud Computing.

Vantagens

- Substituição de CapEx por OpEx.
- Economias de escala.
- Planejamento de capacidade.
- Maior velocidade e agilidade.
- Eliminação de custos com Data Centers.
- Alcance global em minutos.

Conclusão

- Definição de computação em nuvem.
- Principais vantagens da computação em nuvem.

Próxima aula

- Considerações financeiras.
- Considerações de planejamento de capacidade.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 1. Computação em Nuvem

Aula 1.2. Considerações financeiras e de planejamento de capacidade

Prof. Taynan Ferreira

Nesta aula

- Definições de termos financeiros.
- Considerações financeiras quanto ao uso de cloud.
- Considerações de planejamento de capacidade.

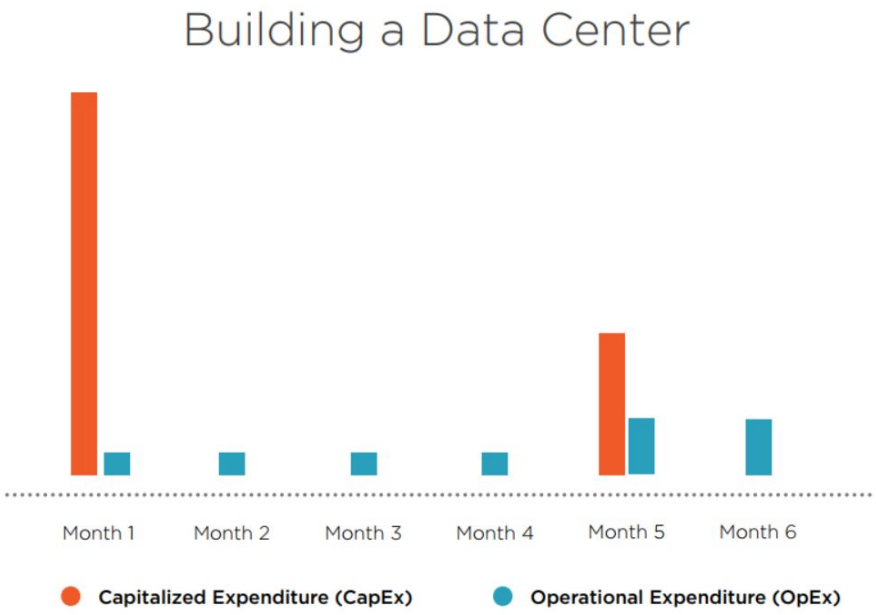
CapEx (Capital Expenditure)

- Investimentos feitos por uma companhia para adquirir ou aprimorar ativos, com o intuito de serem usados a longo prazo.
- Exemplos:
  - Aquisição de propriedades, plantas, edifícios, tecnologia, equipamentos etc.
  - Investimentos CapEx em data centers: aquisição de terreno, construção de edifício, aquisição de servidores etc.

OpEx (Operational Expenditure)

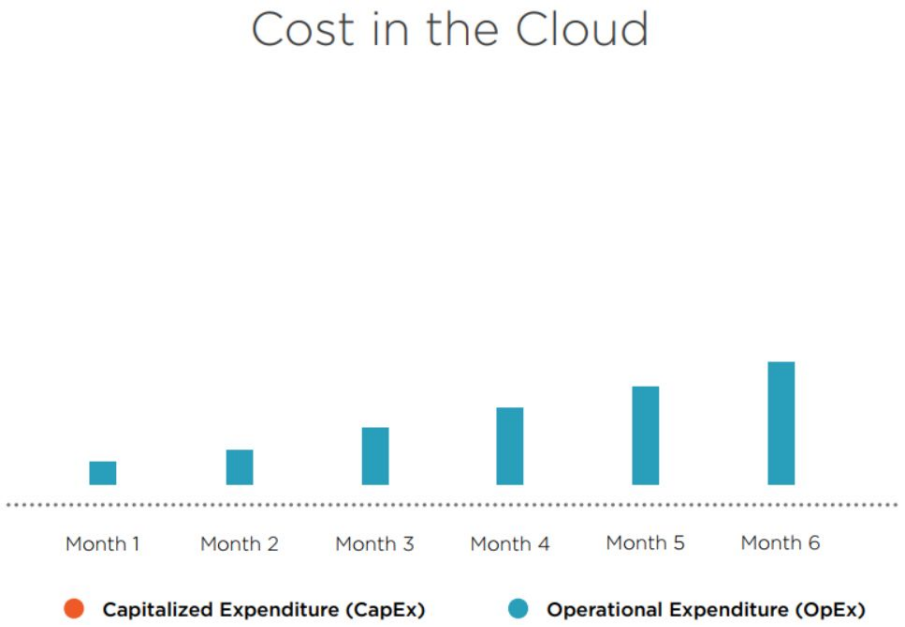
- Custos relacionados à manutenção das atividades do dia a dia das operações de uma empresa.
- Exemplos:
  - Custos com aluguéis, salários, pagamentos de juros etc.
  - Em data centers: custos com energia elétrica, conectividade, água, salário de funcionários etc.

Data Center vs Cloud



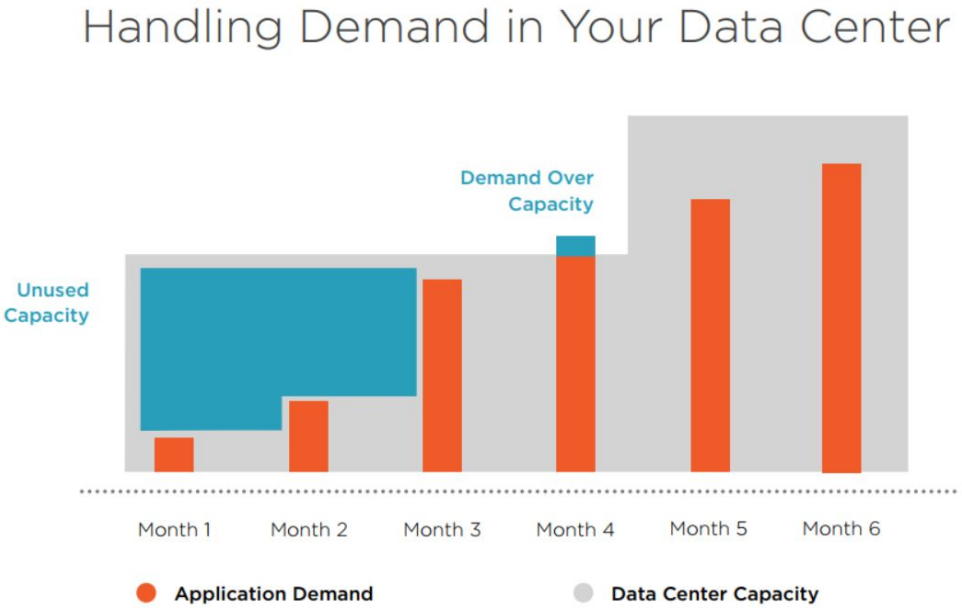
Fonte: Pluralsight: Fundamental Cloud Concepts for AWS (David Tucker).

Data Center vs Cloud



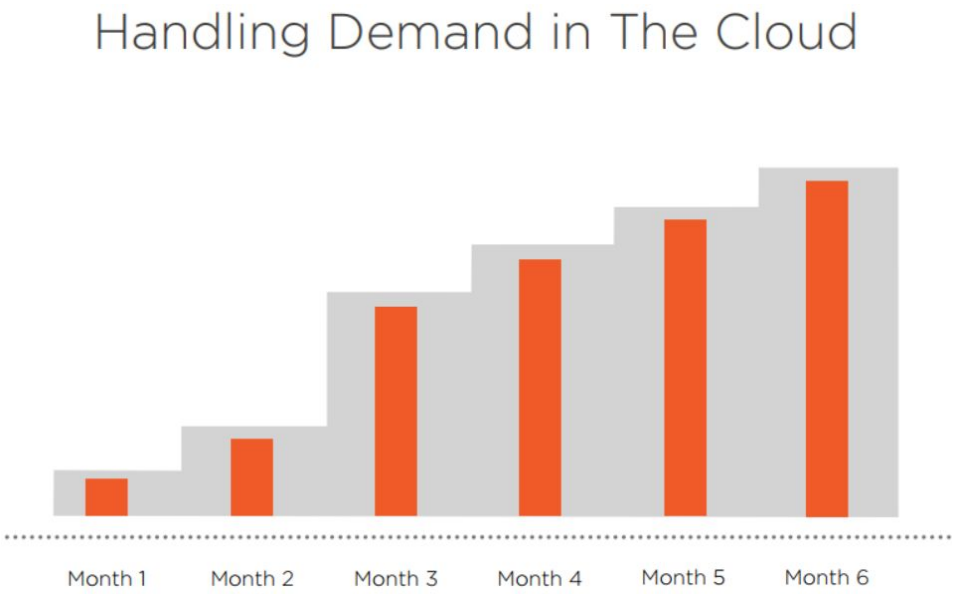
Fonte: Pluralsight: Fundamental Cloud Concepts for AWS. (David Tucker)

Planejamento de capacidade



Fonte: Pluralsight: Fundamental Cloud Concepts for AWS. (David Tucker)

Planejamento de capacidade



Fonte: Pluralsight: Fundamental Cloud Concepts for AWS. (David Tucker)

Conclusão

- Definições de termos financeiros.
- Considerações financeiras quanto ao uso de cloud.
- Considerações de planejamento de capacidade.

Próxima aula

- Conceitos e terminologia básicos.
- Modelos de Deploy na nuvem.
- Modelos de Serviço na nuvem.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 1. Computação em Nuvem

Aula 1.3. Conceitos e terminologia

Prof. Taynan Ferreira

Nesta aula

- Principais conceitos e terminologia.
- Modelos de Deploy na nuvem.
- Modelos de Serviço na nuvem.

Disponibilidade

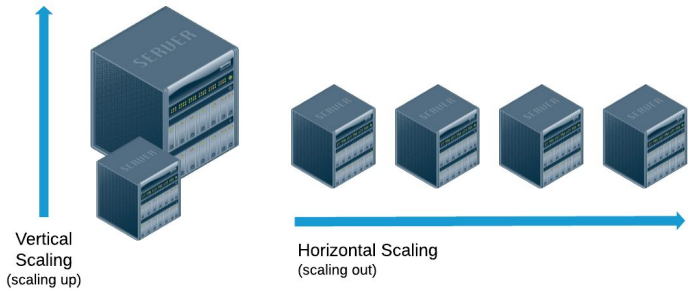
- Capacidade de uma solução de satisfazer as necessidades de negócio para as quais foi projetada. Aplicações de alta disponibilidade são desenhadas de forma que um único ponto de falha não diminua sua capacidade de ser completamente operacional.

Elasticidade

- Capacidade de aquisição de recursos computacionais quando necessário e a liberação desses mesmos recursos quando não mais necessário.

Escalar (*Scaling*)

- Processo de crescimento de recursos computacionais de modo a atender à demanda da aplicação. É possível escalar verticalmente ou horizontalmente.



Fonte: Section.io

Modelos de Deploy

- Público:
  - Aplicação totalmente executada na nuvem pública.
- Privado:
  - Infraestrutura on-premise disponibilizada via virtualização.
- Híbrido:
  - Conexão entre serviços/infraestrutura on-premise e em nuvem.

Modelos de Serviço

- Infrastructure as a Service (IaaS):
  - Provisão da infraestrutura mais fundamental:
    - Armazenamento, rede, sistema operacional, servidores e outros recursos básicos de computação.
- Maior nível de flexibilidade e customização.

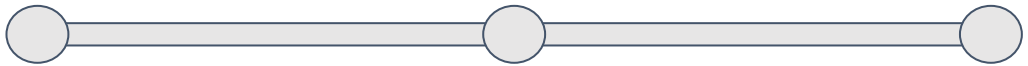
Modelos de Serviço

- Platform as a Service (PaaS):
  - Gerenciamento da infraestrutura a cargo do provedor de nuvem:
    - Planejamento de capacidade.
    - Manutenção de software.
    - Patching etc.
  - Empresas podem se focar unicamente no desenvolvimento, implantação e manutenção de suas aplicações.

Modelos de Serviço

- Software as a Service (IaaS):
  - A aplicação é totalmente gerenciada pelo provedor.
  - Libera do usuário a necessidade de se preocupar com qualquer tipo de manutenção ou planejamento de capacidade.
  - Usuário se preocupa unicamente com o uso do software em si.

Modelos de Serviço



IaaS PaaS SaaS

Maior controle

Maior manutenção

Menor controle

Menor manutenção

Conclusão

- Principais conceitos e terminologia.
- Modelos de Deploy na nuvem.
- Modelos de Serviço na nuvem.

Próxima aula

- Overview do Mercado.
- Principais provedores de cloud computing.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 1. Computação em Nuvem

Aula 1.4. Overview do Mercado de Computação em Nuvem

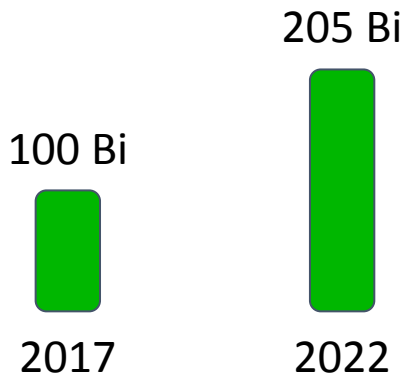
Prof. Taynan Ferreira

Nesta aula

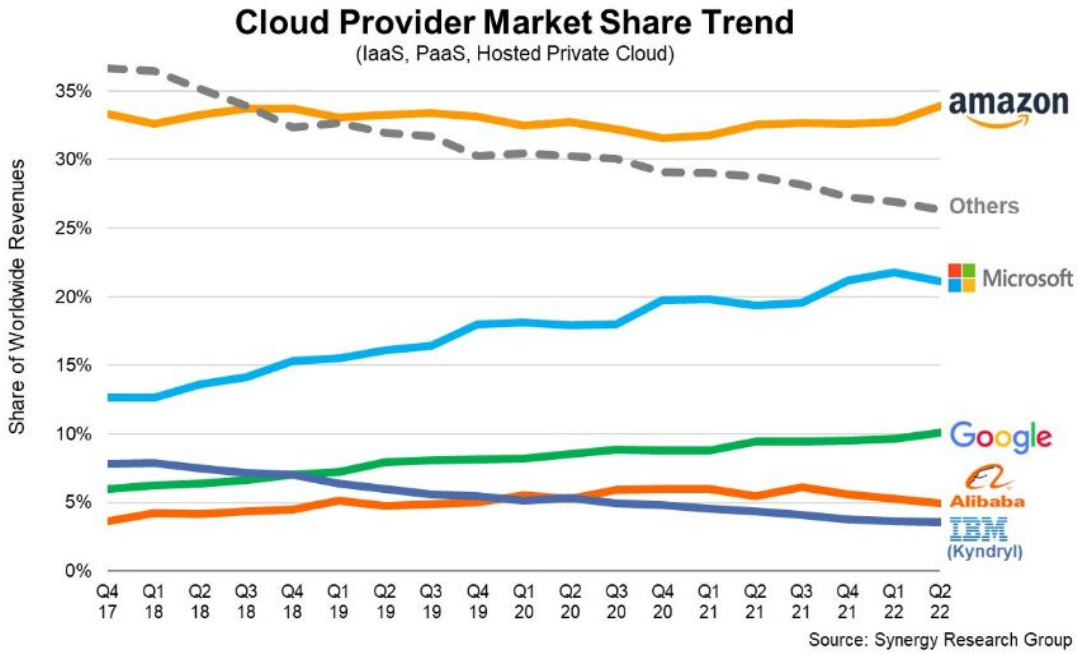
- Tamanho do mercado mundial.
- Principais provedores:
  - Market share.
  - Posicionamento competitivo.

Tamanho do mercado

- Gastos no 1º Semestre de 2022: USD 55 bilhões.
- Receita combinada dos principais cloud providers (USD) em 12 meses.



Market Share



Fonte: [Synergy Research Group](#).

Quadrante mágico de Gartner

Fonte: AWS

Principais provedores

- Amazon: Amazon Web Services (AWS):
  - Pioneiro, foi lançado em 2006.
  - Pontos fortes:
    - Cadeira de Suprimentos e Engenharia.
    - Líder em inovação.
  - Pontos fracos:
    - Complexidade da oferta.

Principais provedores

- Microsoft: Microsoft Azure.
  - Lançado em 2008.
  - Pontos fortes:
    - Conjunto mais amplo de recursos.
    - Destaque em soluções de Big Data.
  - Pontos fracos:
    - Licenciamento e contratação complexos.

Principais provedores

- Google: Google Cloud Platform (GCP).
  - Lançado para o em 2008 como Google App Engine (GAE).
  - Pontos fortes:
    - Velocidade de inovação.
  - Pontos fracos:
    - Satisfação pós-venda.

Conclusão

- Tamanho do mercado mundial.
- Principais provedores:
  - Market share.
  - Posicionamento competitivo.

Próxima aula

- Conhecendo a AWS:
  - Infraestrutura AWS.
  - Custos na AWS.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 2. Conhecendo a AWS

Aula 2.1. Infraestrutura AWS

Prof. Taynan Ferreira

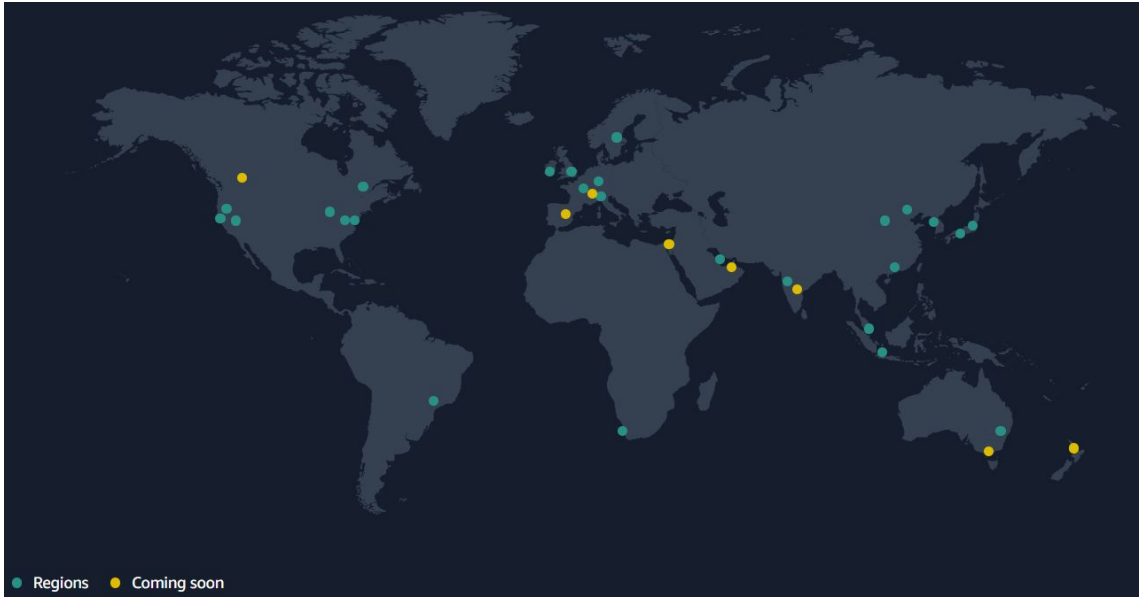
Nesta aula

- Conhecer as Regiões AWS.
- Entender as Zonas de Disponibilidade.

Regiões (AWS Region)

- Cada Região está em uma localização geográfica específica.
- Cada Região possui um cluster de data centers.
- Atualmente (Setembro/2022) a AWS conta com 26 regiões, além de 8 planejadas para serem lançadas em breve.

Distribuições de Regiões



Fonte: AWS

Regiões (AWS Region)

Região	Identificador	Região	Identificador	Região	Identificador
US East (N. Virginia)	us-east-1	Asia Pacific (Sydney)	ap-southeast-2	South America (São Paulo)	sa-east-1
US East (Ohio)	us-east-2	Asia Pacific (Tokyo)	ap-northeast-1	Canada (Central)	ca-central-1
US West (N. California)	us-west-1	Canada (Central)	ca-central-1	Europe (Frankfurt)	eu-central-1
US West (Oregon)	us-west-2	Europe (Frankfurt)	eu-central-1	Europe (Ireland)	eu-west-1
Africa (Cape Town)	af-south-1	Europe (Ireland)	eu-west-1	Europe (London)	eu-west-2
Asia Pacific (Hong Kong)	ap-east-1	Europe (London)	eu-west-2	Europe (Milan)	eu-south-1
Asia Pacific (Jakarta)	ap-southeast-3	Europe (Milan)	eu-south-1	Europe (Paris)	eu-west-3
Asia Pacific (Mumbai)	ap-south-1	Europe (Paris)	eu-west-3	Europe (Stockholm)	eu-north-1
Asia Pacific (Osaka)	ap-northeast-3	Europe (Stockholm)	eu-north-1	Middle East (Bahrain)	me-south-1
Asia Pacific (Seoul)	ap-northeast-2	Middle East (Bahrain)	me-south-1	Middle East (UAE)	me-central-1
Asia Pacific (Singapore)	ap-southeast-1	Middle East (UAE)	me-central-1	South America (São Paulo)	sa-east-1

Zonas de Disponibilidade (AZs)

- Consistem de um ou mais data centers.
- Múltiplas Zonas de Disponibilidade estão incluídas em cada Região AWS.
- Estão localizadas na área geográfica da Região AWS.
- Possuem redundância em termos de energia elétrica, rede e conectividade.
- Atualmente a infraestrutura AWS conta com 87 Zonas de Disponibilidade.

Nomenclatura Regions/AZs

Fonte: Pluralsight: Fundamental Cloud Concepts for AWS (David Tucker).

Conclusão

- Conhecer as Regiões AWS.
- Entender as Zonas de Disponibilidade.

Próxima aula

- Entender como funcionam os custos na AWS.
- Principais serviços de acompanhamento de custos.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 2. Conhecendo a AWS

Aula 2.2. Custos na AWS

Prof. Taynan Ferreira

Nesta aula

- Conhecer os drivers de custo na AWS.
- Aprender sobre ofertas gratuitas disponíveis.
- Conhecer serviços de acompanhamento e gerenciamento de custos.

Drivers de Custo

- Recursos de computação.
- Armazenamento.
- Tráfego de saída.

Ofertas gratuitas (*free tier*)

- 12 meses gratuitos.
- Testes gratuitos.
- Para sempre gratuitos.

Acompanhamento de custos

- AWS Cost Explorer.
- AWS Budget.

Conclusão

- ? Conheceu os drivers de custo da AWS.
- ? Aprendeu sobre ofertas gratuitas disponíveis.
- ? Conheceu serviços de acompanhamento e gerenciamento de custos.

Próxima aula

- Interagir com a AWS.
- Conhecer o AWS Cost Explorer e AWS Budget.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 3. Overview Soluções AWS

Aula 3.1. Amazon SageMaker

Prof. Taynan Ferreira

Nesta aula

- O que é o Amazon SageMaker?
- Conhecer as principais funcionalidades.

Amazon SageMaker

- Serviço de Machine Learning totalmente gerenciado.
- Permite o treino e implantação de modelos.
- Provê instância integrada de notebook Jupyter.

Funcionalidades

- SageMaker Studio:
  - Ambiente integrado para construção, treino, implantação e análise de modelos.
- SageMaker Serverless Endpoints:
  - Endpoint serverless para inferência do modelo.
  - Capacidade escala automaticamente.
  - Evita a necessidade de definição de tipos de instâncias ou políticas de escalabilidade.

Funcionalidades

- SageMaker Data Wrangler:
  - Importação, análise e preparação de dados no SageMaker Studio.
  - Pré-processamento e engenharia de atributos (*Feature Engineering*).
- Feature Store:
  - Catálogo centralizado de features e metadados.
  - Online Store: inferência em tempo real de baixa latência.
  - Offline Store: treinamento e inferência batch.

Funcionalidades

- SageMaker Autopilot:
  - Automatic Machine Learning (AutoML).
  - Automatiza processo de seleção de algoritmo, preparação de dados, treino e tuning.
- SageMaker Model Monitor:
  - Monitoramento de modelos produtivos.
  - Detecção de data drift e problemas de qualidade.
- Batch Transform:
  - Inferências batch.

Conclusão

- Conheceu o Amazon SageMaker.
- Visão geral das principais funcionalidades.

Próxima aula

- Demonstração: títulos públicos com Amazon SageMaker.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 3. Overview Soluções AWS

Aula 3.3. Amazon S3

Prof. Taynan Ferreira

Nesta aula

- O que é o Amazon S3?
- Conhecer as principais características.

Amazon S3

- Serviço de armazenamento de objetos.
- Alta escalabilidade, disponibilidade, segurança e desempenho.
- Vasto espectro de casos de uso:
  - Data Lake
  - Websites
  - Aplicações mobile
  - Backup e Restauração
  - Big Data Analytics

Características

- Distintas classes de armazenamento:
  - S3 Standard.
  - S3 Glacier.
  - S3 Intelligent-Tiering.

- Versionamento:
  - Múltiplas versões do objeto no mesmo bucket.
  - Restauração de objetos removidos/modificados acidentalmente.

- Diversos outros recursos:
  - Controle de acesso, processamento de dados, logging, etc.

Overview Amazon S3

- Objetos são armazenados em Buckets.
  - Objetos:
    - Arquivo.
    - Metadados.
  - Bucket:
    - Container para objetos.

Overview Amazon S3

- Exemplo:
  - Bucket: DOC-EXAMPLE-BUCKET
  - Objeto: photos/puppy.jpg
  - Endereço URL: https://DOC-EXAMPLE-BUCKET.s3.us-west-2.amazonaws.com/photos/puppy.jpg

Conclusão

- Conheceu o Amazon S3.
- Visão geral das principais características.

Próxima aula

- Demonstração: títulos públicos com Amazon SageMaker e Amazon S3.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 3. Overview Soluções AWS

Aula 3.5. AWS Glue

Prof. Taynan Ferreira

Nesta aula

- O que é o AWS Glue.
- Conhecer as principais características.

AWS Glue

- Serviço de integração de dados serverless:
  - Descobrir, preparar, mover e integrar dados de múltiplas fontes.
- Permite conexão com mais de 70 fontes de dados.
- Gerenciamento de catálogo de dados centralizado.
- Consulta de dados catalogados via:
  - Amazon Athena.
  - Amazon EMR.
  - Amazon Redshift Spectrum.

Funcionalidades

- Descobrir e organizar dados:
  - Inferência de schema e catalogação de metadados automatizado.
  - Conexão a múltiplas fontes de dados.

- Transformar, preparar e limpar dados para análise:
  - Construção de pipelines de ETLs complexos com agendamento.
  - Limpeza e transformação de dados em streaming.

- Construir e monitorar pipelines de dados:
  - Infraestrutura escala automaticamente.
  - Tarefas automatizados via *triggers* baseados em eventos.

Terminologia AWS Glue

- AWS Glue Data Catalog:
  - Repositório de metadados.
  - Contém definições de tabelas, tarefas, entre outros..
  - Cada conta AWS possui um Data Catalog por Região AWS

- Crawler:
  - Programa que conecta a um repositório de dados, processa uma lista de classificadores para inferência de schema e cria o metadados no AWS Glue Data Catalog.

Conclusão

- Conheceu o Amazon Glue.
- Visão geral das principais características.

Próxima aula

- Demonstração: títulos públicos com Amazon SageMaker e AWS Glue.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 4. Introdução ao SQL

Aula 4.1. Introdução ao SQL

Prof. Taynan Ferreira

Nesta aula

- Conhecer instruções básicas de SQL.
- Trabalhar exemplos práticos de DDL, DML e DQL.

Linguagem de Definição de Dados (DDL)

- Utilizado para definir, alterar ou excluir objetos do banco de dados:
  - Criar: CREATE.
  - Alterar: ALTER.
  - Excluir: DROP.

Linguagem de Definição de Dados (DDL)

- Exemplo de criação de tabela:

```
CREATE TABLE tesouro_direto
 (data_negociacao DATE,
  taxa_compra DECIMAL(8, 2),
  taxa_venda DECIMAL(8, 2),
  pu_compra DECIMAL(8, 2),
  pu_venda DECIMAL(8, 2)
 );
```

Linguagem de Definição de Dados (DDL)

- Exemplo de exclusão de tabela:

```
DROP TABLE tesouro_direto;
```

Linguagem de Manipulação de Dados (DML)

- Utilizado para adicionar, atualizar ou excluir registros em tabelas de banco de dados:
  - Adicionar: INSERT.
  - Atualizar: UPDATE.
  - Excluir: DELETE.

Linguagem de Manipulação de Dados (DML)

- Exemplo de inserção de registro:

```
INSERT INTO "xpe"."tesouro_prefixado"(
"data base", "taxa compra manha", "taxa venda manha", "pu compra manha", "pu
venda manha", "pu base manha")
VALUES ('2019-02-01', 0.02, 0.06, 9926.99, 9902.94, 9900.45);
```

- Exemplo de exclusão de registro:

```
DELETE FROM "xpe"."tesouro_prefixado"
WHERE “data base” = '2019-02-01';
```

Linguagem de Consulta de Dados (DQL)

- Utilizado para consultar dados, realizando filtros, cálculos, agrupamentos, etc. Cláusulas de DQL:
  - SELECT: colunas a serem retornadas.
  - FROM: tabelas consultadas.
  - WHERE: filtrar dados.
  - GROUP BY: agrupar linhas.
  - HAVING: filtra grupo de linhas.
  - ORDER BY: ordenação de resultado final.

Linguagem de Consulta de Dados (DQL)

- Exemplo de *query:*

```
SELECT
    MIN("taxa compra manha") as taxa_minima,
    MAX("taxa compra manha") as taxa_maxima
FROM "xpe"."titulos_publicos"
WHERE "data base" < '2021-12-31'
```

Conclusão

- Conhecer instruções básicas de SQL.
- Trabalhar exemplos práticos de DDL, DML e DQL.

Próxima aula

- Conhecer o Amazon Athena
- Aprender sobre principais características desse serviço.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 4. Introdução ao SQL

Aula 4.2. Amazon Athena

Prof. Taynan Ferreira

Nesta aula

- O que é Amazon Athena?
- Conhecer as principais características.

Amazon Athena

- Serviço de consulta interativa a dados no Amazon S3 via SQL.
- Serverless:
  - Não exige configuração e manutenção de infraestrutura.
  - Pagamento apenas por *queries* executadas.
- Escala automaticamente:
  - Queries executadas em paralelo.
  - Resultados rápidos mesmo com datasets grandes e *queries* complexas.

Características

- Análise de dados em S3:
  - Dados estruturados e não estruturados.
  - Exemplos de objetos: CSV, JSON, Apache Parquet.
- Integra com o AWS Glue Data Catalog.
- Caso de uso:
  - Execução de *queries* ad hoc em dados armazenados no S3.

Conclusão

- Conheceu o Amazon Athena
- Visão geral das principais características.

Próxima aula

- Demonstração: títulos públicos com Amazon S3, AWS Glue e Amazon Athena.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 4. Introdução ao SQL

Aula 4.4. Explicação do Trabalho Prático

Prof. Taynan Ferreira

Nesta aula

- Rever o conteúdo coberto na 1ª parte do módulo.
- Apresentar o Trabalho Prático.

Conteúdo coberto

- Computação em nuvem:
  - Definição, vantagens e considerações financeiras.
  - Conceitos e terminologias básicas.
  - Overview do Mercado.
- AWS:
  - Infraestrutura AWS.
  - Custos na AWS.
  - Overview de soluções AWS para Big Data & Analytics (S3, Glue, SageMaker e Athena).

Conteúdo coberto

- SQL:
  - DDL, DML e DQL.
  - Principais instruções SQL.
- Demonstrações Práticas:
  - Uso de Amazon Sagemaker, S3, Glue e Athena.
  - Coleta e armazenamento de dados de Títulos Públicos.

Objetivos Trabalho Prático

- Reforçar conteúdo coberto no módulo.
- Desenvolver experiência prática com as ferramentas apresentadas.

Trabalho Prático

- Coletar, armazenar e consultar dados do Tesouro Direto:
  - Coletar dados de todos os títulos públicos disponíveis utilizando API Tesouro Transparente.
  - Salvar dados coletados no Amazon S3.
  - Construir metadados no Amazon Glue Data Catalog através do Glue Crawler.
  - Consultar dados armazenados utilizando SQL e Amazon Athena.
  - Consultas sobre Tesouro IPCA+ 2035 e Tesouro Prefixado 2025.

Trabalho Prático

- Orientações Gerais
  - Validar carregamento dos dados.
  - Limitar consulta aos dados até o fim de 2021.
  - Eliminar recursos após o trabalho prático.

Conclusão

- Rever o conteúdo coberto na 1ª parte do módulo.
- Apresentar o Trabalho Prático.

Próxima aula

- Fundamentos de Engenharia de Dados:
  - Tipos de Dados.
  - Tipos de Processamento de Dados.
  - Principais Modelos de Dados.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 5. Fundamentos de Engenharia de Dados

Aula 5.1. Tipos de Dados e Modelos de Dados

Prof. Taynan Ferreira

Nesta aula

- Conhecer os tipos de dados.
- Entender os tipos de processamento de dados.
- Estudar os principais modelos de dados.

Tipos de Dados

- Dados estruturados:
  - Estrutura bem-definida.
  - Diferença entre cada registro: conteúdo.
  - Facilmente organizados em formato tabular.

Tipos de Dados

- Dados não estruturados:
  - Não possuem estrutura fixa e predefinida.
  - Exemplos: texto, imagem, áudio, vídeo etc.
  - Maior parte dos dados.

Tipos de Processamento de Dados

- Online Transaction Processing (OLTP):
  - Orientado a transações (ou eventos).
  - Origem: interação entre usuários, empresas e aplicações.
  - Registros inseridos/atualizados individualmente.

Tipos de Processamento de Dados

- Online Analytical Processing (OLAP):
  - Surgimento atrelado ao nascimento de Data Analytics.
  - Análise e extração de valor de grande quantidade de dados históricos.
  - Utilizado por equipes internas para a geração de relatório e tomada de decisões.

Tipos de Processamento de Dados

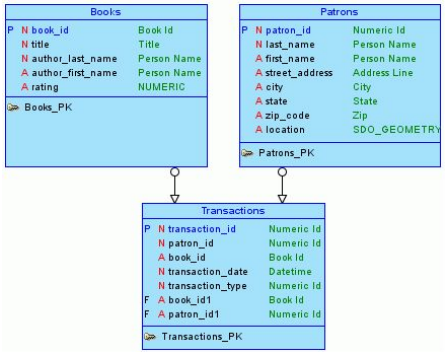
Propriedade	OLTP	OLAP
Padrão de leitura	Pequena quantidade de registros por consulta	Agregações sobre grande quantidade de dados
Padrão de escrita	Baixa latência com entradas por usuário	Carga (ETL) ou Streaming de eventos
Utilizado primariamente por	Usuário final	Analistas internos, como suporte à decisão
Dado representa	Estado atual do dado	Histórico de eventos ao longo do tempo

Modelos de Dados

- Modelo Relacional.
- Modelo orientado a Documentos.
- Modelo orientado a Grafos.

Modelo Relacional

- Modelo para o qual foi desenvolvida a linguagem SQL.
- Informações são organizadas em relações (tabelas em SQL).
- Cada relação é uma coleção não ordenada de tuplas (linhas em SQL).



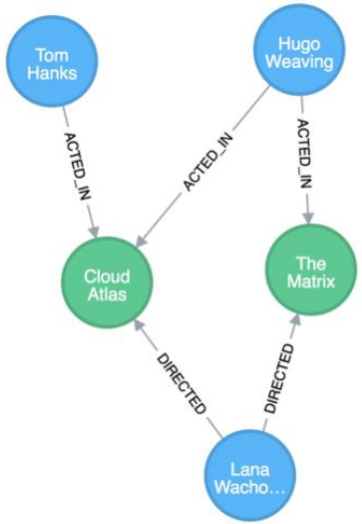
Fonte: [Oracle Tutorial on Data Modeling](#).

Orientado a Documentos

- Maior flexibilidade de estrutura (schema on read).
- Melhor desempenho em alguns cenários (localidade do dado).
- Reflete melhor os modelos de dados utilizados em programação orientada a objetos (redução de impedance mismatch).
- Dado armazenado em documento (e.g. JSON ou XML).

Orientado a Grafos

- Registros organizados em:
  - Vértices (nós ou entidades).
  - Arestas (relações ou arcos).
- Ideal para relações n-n (many-to-many):
  - Redes sociais.
  - Links entre páginas Web.
  - Malha viária.



Fonte: Neo4j Developer Guide.

Conclusão

- Conhecer os tipos de dados.
- Entender os tipos de processamento de dados.
- Estudar os principais modelos de dados.

Próxima aula

- Modelo de dados: caso de uso.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 5. Fundamentos de Engenharia de Dados

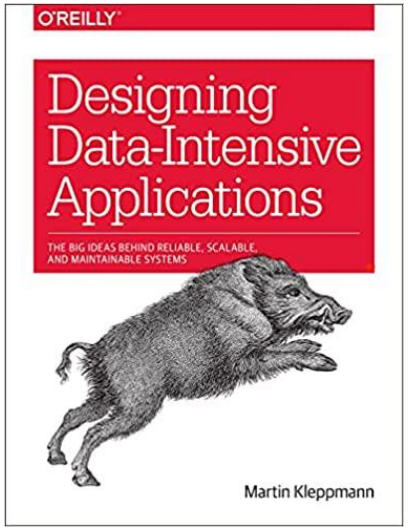
Aula 5.2. Modelos de Dados: caso de uso

Prof. Taynan Ferreira

Nesta aula

- Caso de uso em Modelos de Dados.

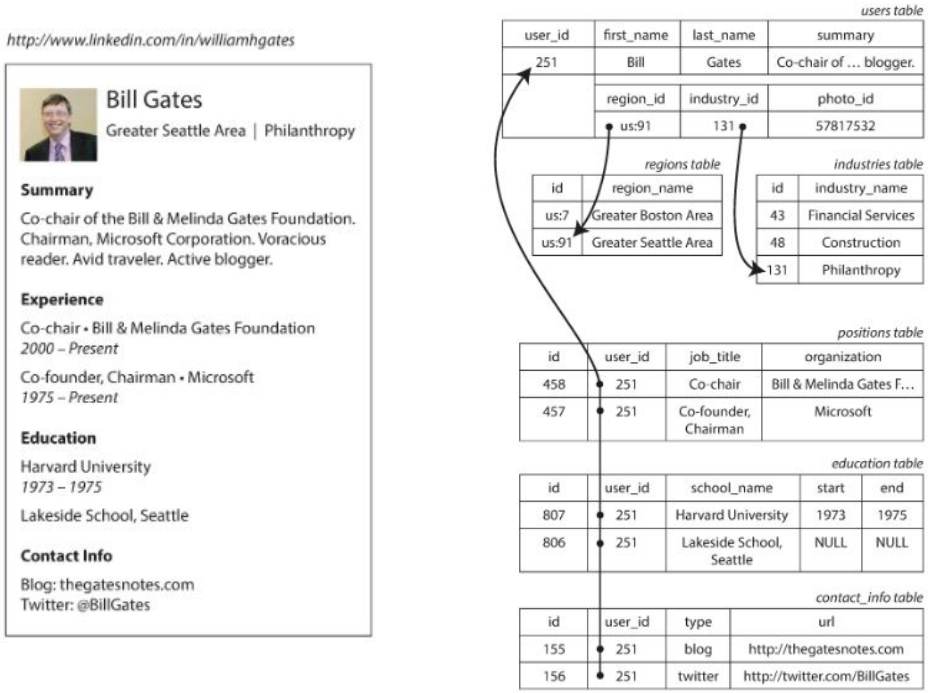
Modelos de Dados: exemplo



http://www.linkedin.com/in/williamhgates

Fonte: Designing Data-Intensive Applications.

Modelos de Dados: exemplo



Fonte: Designing Data-Intensive Applications.

Modelos de Dados: exemplo

http://www.linkedin.com/in/williamhgates

Bill Gates

Greater Seattle Area | Philanthropy

Summary

Co-chair of the Bill & Melinda Gates Foundation. Chairman, Microsoft Corporation. Voracious reader. Avid traveler. Active blogger.

Experience

Co-chair • Bill & Melinda Gates Foundation
2000 – Present

Co-founder, Chairman • Microsoft
1975 – Present

Education

Harvard University
1973 – 1975

Lakeside School, Seattle

Contact Info

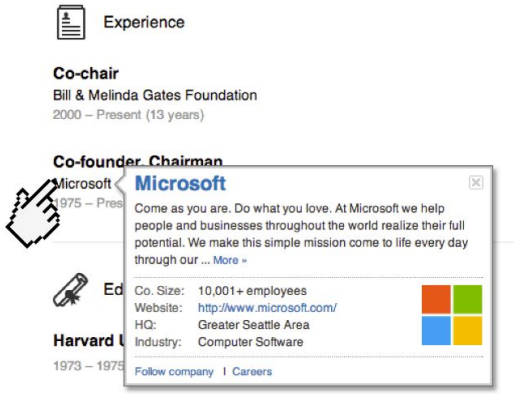
Blog: thegatesnotes.com
Twitter: @BillGates


```
{
  "user_id":     251,
  "first_name":  "Bill",
  "last_name":   "Gates",
  "summary":     "Co-chair of the Bill & Melinda Gates... Active blogger.",
  "region_id":   "us:91",
  "industry_id": 131,
  "photo_url":   "/p/7/000/253/05b/308dd6e.jpg",
  "positions": [
    {"job_title": "Co-chair", "organization": "Bill & Melinda Gates Foundation"},
    {"job_title": "Co-founder, Chairman", "organization": "Microsoft"}
  ],
  "education": [
    {"school_name": "Harvard University",       "start": 1973, "end": 1975},
    {"school_name": "Lakeside School, Seattle", "start": null, "end": null}
  ],
  "contact_info": {
    "blog":    "https://www.gatesnotes.com/",
    "twitter": "https://twitter.com/BillGates"
  }
}
```


Fonte: Designing Data-Intensive Applications.

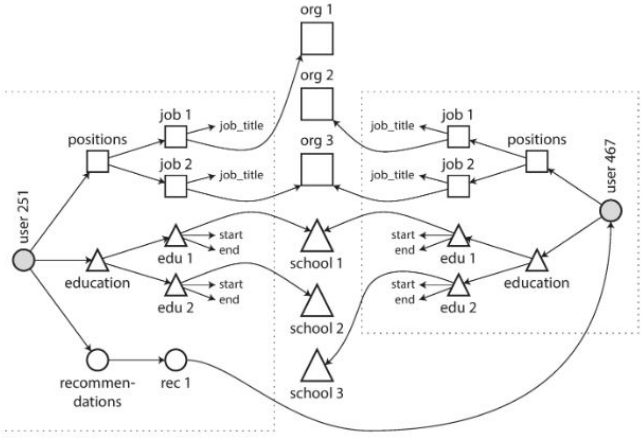
Modelos de Dados: exemplo

- Novas features:
  - Empresas e universidades como entidades.
  - Recomendações.



Fonte: Designing Data-Intensive Applications.

Modelos de Dados: exemplo



Fonte: Designing Data-Intensive Applications.

- Que modelo de dados utilizar?
  - Relacional.
  - Orientado a Grafos.

Conclusão

? Caso de uso em Modelos de Dados.

Próxima aula

- Pipeline de Ciência de Dados.
- Framework CRISP-DM.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 6. Pipeline de Ciência de Dados

Aula 6.1. Pipeline de Ciência de Dados

Prof. Taynan Ferreira

Nesta aula

- Conhecer o framework CRISP-DM.
- Entender cada etapa do pipeline de Ciência de Dados.

Objetivo do Framework

- Catalogar e guiar processo.
- Identificar boas práticas.
- Evitar erros comuns.
- Sistematizar: Dado → Conhecimento.

CRISP-DM

Cross Industry Standard Process

for Data Mining

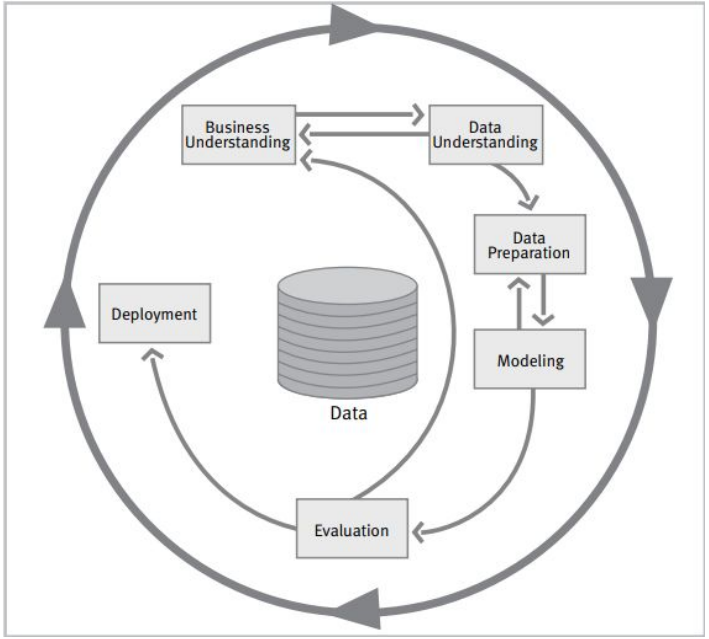
“Processo Padrão Inter-Indústrias

para Mineração de Dados”

Data Mining?

Fonte: Google Trends. Visitado em 01/10/2022

CRISP-DM



Fonte: CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories

- Entendimento do negócio.
- Entendimento dos dados.
- Preparação dos dados.
- Modelagem.
- Avaliação.
- Implantação.

Entendimento do negócio

- Determinar objetivos de negócio.
- Avaliar situação (recursos, riscos, custos etc.).
- Determinar critérios de sucesso.
- Planejar o projeto.

Entendimento dos dados

- Coletar dados iniciais.
- Descrever dados.
- Explorar dados.
- Verificar qualidade dos dados.

Preparação dos dados

- Selecionar dados.
- Limpeza de dados.
- Construir e integrar os dados.

Modelagem

- Selecionar técnicas de modelagem.
- Construir o modelo.
- Avaliar o modelo.

Avaliação

- Avaliar resultados.
- Determinar os próximos passos.

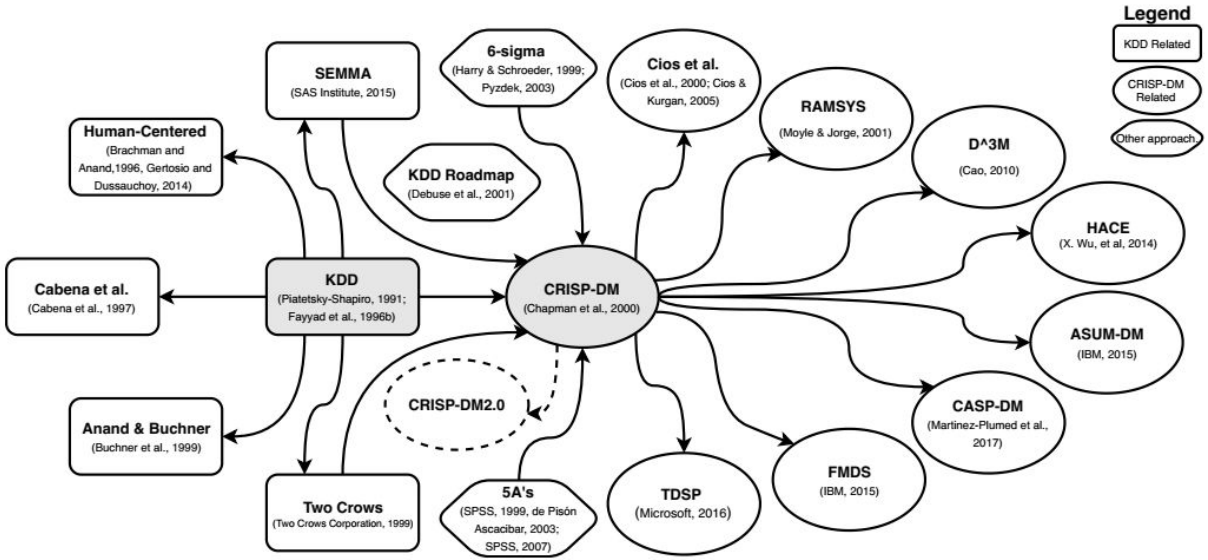
Implantação

- Planejar a implantação.
- Planejar o monitoramento e manutenção.

Evolução da disciplina

- Data Mining (início dos anos 2000):
  - Orientado pelos objetivos de negócio.
  - Foco no processo.
- Data Science (anos 2020):
  - Orientado pelos dados.
  - Exploratório.

Outros Frameworks



Fonte: CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories

Conclusão

- Conhecer o framework CRISP-DM.
- Entender cada etapa do pipeline de Ciência de Dados.

Próxima aula

- Processamento de Linguagem Natural.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 7. Processamento de Linguagem Natural

Aula 7.1. Processamento de Linguagem Natural

Prof. Taynan Ferreira

Nesta aula

- Definição Processamento de Linguagem Natural (NLP).
- Aplicações de NLP.
- Técnicas de pré-processamento de Linguagem Natural.

Definição

- “Processamento de Linguagem Natural é o conjunto de técnicas para tornar a linguagem humana acessível aos computadores”

Jacob Eisenstein

Aplicações

- Classificação de texto.
- Recuperação de texto (Information Retrieval).
- Sumarização de texto.
- Tradução de texto (Machine Translation).

Pré-processamento

- Facilita a transformação de documentos em representações matemáticas relevantes e passíveis de serem tratadas computacionalmente (vetores, matrizes etc.).
- Podemos agrupar em:
  - Tokenização.
  - Normalização.
  - Anotação.

Tokenização

- Textos → Unidades de processamento:

“At eight o'clock on Thursday morning... Arthur didn't feel very good.”

['At', 'eight', "o'clock", 'on', 'Thursday', 'morning',

'Arthur', 'did', "n't", 'feel', 'very', 'good', '.']

Tokenização

- Token: unidade semântica de interesse.
- Tipo: conjunto de tokens idênticos.
- Termo: tipo de fato considerado na análise/modelagem.

Tokenização

“*to sleep, perchance to dream*” (Hamlet, Shakespeare)

5 tokens: ['to', 'sleep', 'perchance', 'to', 'dream']

4 tipos: ['to', 'sleep', 'perchance', 'dream']

3 termos: ['sleep', 'perchance', 'dream']

Tokenização

- É uma tarefa simples? Simplesmente separar por espaços?
- Como tokenizar:
  - Rio de Janeiro.
  - Fernando Pessoa.
  - Fim de semana.

Normalização

Tokens com grafias distintas (mas mesmo significado)

Tokens com a mesma grafia

Normalização: exemplos

- Case folding:
  - “Literatura”, “literatura”, “LITERATURA” → “literatura”
- Stemização (Stemming):
  - “informa”, “informação”, “informática” → “inform”
- Lematização (Lemmatization):
  - “corria”, “correria”, “correu” → “correr”

Anotação

- Inverso da Normalização.
- Normalização: iguala tokens de grafias distintas.
- Anotação: diferencia tokens de grafia idêntica.

Anotação: exemplo

- fly: voar (verbo) ou mosca (substantivo)
- fly → fly/VB ou fly/NN

Conclusão

- Definição Processamento de Linguagem Natural (NLP).
- Aplicações de NLP.
- Técnicas de pré-processamento de Linguagem Natural.

Próxima aula

- Motivação do Desafio: algoritmo para leitura de atas de bancos centrais.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 7. Processamento de Linguagem Natural

Aula 7.2. Motivação do Desafio

Prof. Taynan Ferreira

Nesta aula

- Conhecer aplicação real do uso de NLP para Renda Fixa.
- Ter contato com a motivação para o desafio.

Motivação

Fonte: Itaú BBA

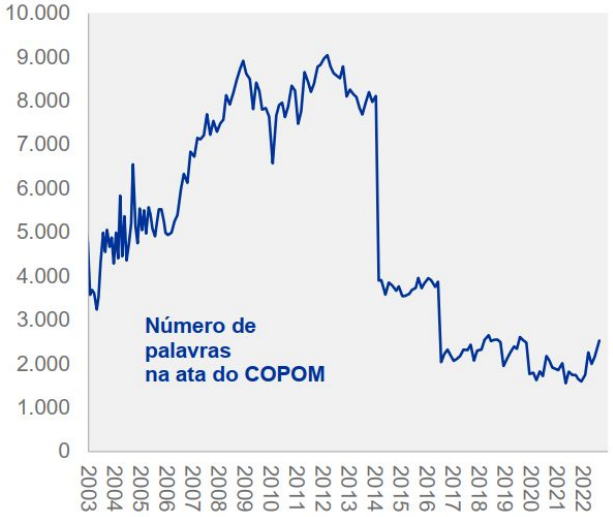
Objetivo

- Ferramenta de interpretação das atas do Copom.
- Aplicação de NLP para mitigar vieses.
- Auxílio como instrumento preditivo das decisões do COPOM.

Metodologia

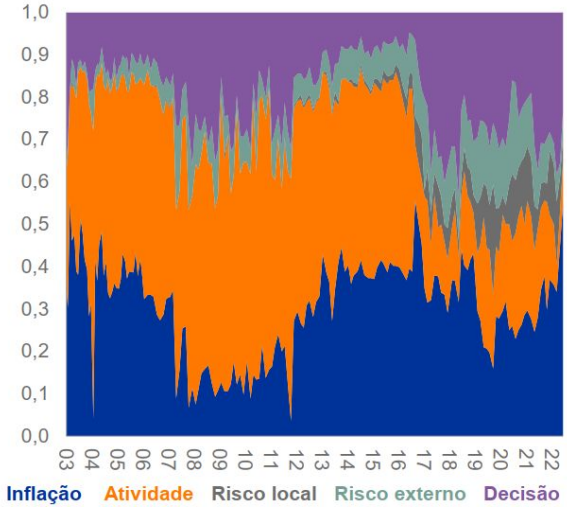
- Análise exploratória

Número de palavras



Fonte: Itaú BBA

Principal assunto



Fonte: Itaú BBA

Metodologia

- Atas do período de 2016 a 2022
  - 14 atas: 585 sentenças únicas.
  - Amostragem: manutenção, corte e alta.

Fonte: Itaú BBA

Metodologia

- Classificação das sentenças das atas em
  - Dove: indicação de baixa.
  - Neutro: indicação de manutenção.
  - Hawk: indicação de alta.

Metodologia

- Trigramas para visualização e conferência

Preditas como dove

Preditas como hawk

Fonte: Itaú BBA

Resultados

- Indicador com boa precedência temporal sobre ciclo de juros

Conclusões

- Conhecer aplicação real do uso de NLP para Renda Fixa.
- Ter contato com a motivação para o desafio.

Próxima aula

- Demonstração: aplicando NLP às comunicações do Banco Central.

Faculdade XPe

Coleta e Armazenamento de Dados de Renda Fixa

Capítulo 7. Processamento de Linguagem Natural

Aula 7.4. Revisão e Apresentação do Desafio

Prof. Taynan Ferreira

Nesta aula

- Rever o conteúdo coberto na 2ª parte do módulo.
- Apresentar o Desafio Prático.

Conteúdo coberto

- Fundamentos de Engenharia de Dados.
  - Tipos de Dados
  - Tipos de Processamento de Dados
  - Modelos de Dados
- Pipeline de Ciência de Dados.
  - Principais etapas
  - Framework CRISP-DM

Conteúdo coberto

- Processamento de Linguagem Natural (NLP).
  - Definição
  - Aplicações de NLP
  - Técnicas de Pré-processamento
- Case real de NLP aplicado à Renda Fixa.
  - Visão geral do artigo
  - Demonstração com atas do COPOM

Objetivos Desafio Prático

- Reforçar conteúdo aprendido no módulo.
- Desenvolver experiência prática com as ferramentas apresentadas.

Desafio Prático

- Coletar, armazenar e consultar atas do COPOM
  - Atas entre os anos de 2019 e 2021
  - Transformar PDFs em TXTs
  - Conteúdo: 3ª página em diante
  - Eliminar Stop Words

- Analisar atas e responder perguntas:
  - Quantidade de tokens únicos
  - Média de palavras distintas
  - Termos mais recorrentes

Desafio Prático

- Orientações importantes
  - Limitar coleta de dados aos anos de 2019 e 2021
  - Validar carregamento dos dados
  - Eliminar recursos após o fim do desafio prático

Conclusões

- Revisão do conteúdo coberto na 2ª parte do módulo.
- Apresentação do Desafio Prático.

Próxima aula

- Segunda aula interativa.