

Arquiteturas em Nuvem para Data Warehouses

Comparação entre AWS, Azure e GCP



Sistemas de informação II – 2024/2025
Licenciatura em Engenharia Informática

Bruno Martins / 2022147149

João Rosa / 2022131973

Índice

• Resumo	3
• Introdução	4
• Contextualização	5
• Estado da Arte Genérica sobre Data Warehousing	6
• AWS – Amazon Redshift	7
• Microsoft Azure – Synapse Analytics	8
• Google Cloud Platform – BigQuery	9
• Comparação Geral: Custo, Escalabilidade e Performance	10
• Conclusão	11
• Referências	12

Resumo

- **Objetivo:** Comparar as arquiteturas de Data Warehouses nas três principais plataformas de computação em nuvem:
 - **Amazon Web Services (AWS) – Amazon Redshift**
 - **Microsoft Azure – Synapse Analytics**
 - **Google Cloud Platform (GCP) – BigQuery**
- **Motivação:** Migração de Data Warehouses para a nuvem oferece vantagens como:
 - Elasticidade
 - Redução de custos
 - Acessibilidade global
- **Conteúdo:** Análise de desempenho, custo, escalabilidade e segurança, além de estudos de caso reais.

Introdução

Importância dos Data Warehouses (DWs):

- Soluções que suportam a gestão de grandes volumes de dados e auxiliam na tomada de decisões informadas.
- **Desafios dos DWs locais (on-premise):**
 - **Alto custo:** Investimentos em hardware, manutenção e gestão especializada.
 - **Falta de escalabilidade:** Limitação de capacidade de processamento e armazenamento.
 - **Manutenção complexa:** Necessidade constante de atualizações e intervenções técnicas.
- **Benefícios da migração para a nuvem:**
 - **Flexibilidade:** Escalabilidade elástica e dinamismo na alocação de recursos.
 - **Redução de custos:** Modelo pay-as-you-go.
 - **Acesso remoto:** Facilita a colaboração e operações globais.



Contextualização

- **Evolução dos Data Warehouses:**
 - Introduzido no final dos anos 80 com o objetivo de centralizar dados provenientes de várias fontes, facilitando relatórios e análises.
 - Inicialmente implementados em servidores locais, com desafios significativos, como:
 - **Custos elevados:** Manter servidores físicos e infraestrutura interna.
 - **Falta de escalabilidade:** Capacidade limitada pela infraestrutura física.
 - **Manutenção complexa:** Exigência de equipes técnicas especializadas.
- **Revolução da Nuvem:**
 - **Elasticidade:** Escalabilidade de acordo com a demanda.
 - **Redução de custos:** Empresas pagam apenas pelo uso efetivo.
 - **Acesso global:** Colaboração facilitada e acessibilidade de qualquer local.
 - **Integração com tecnologias avançadas:** Machine Learning e análises preditivas.

Estado da Arte Genérica sobre Data Warehousing

- **Definição de Data Warehouse (DW):**
 - Sistema que centraliza e integra dados de diversas fontes, facilitando a análise e tomada de decisões.
- **Arquitetura tradicional de um DW:**
 1. **ETL** (Extração, Transformação e Carregamento): Processo de preparar dados para análise.
 2. **Armazenamento centralizado:** Estrutura de dados organizada, geralmente em esquemas estrela ou floco de neve.
 3. **Ferramentas de análise:** Interfaces e ferramentas de Business Intelligence (BI) para explorar os dados.
- **Desafios atuais:**
 - **Crescimento exponencial de dados:** Aumento dos custos associados ao armazenamento e processamento.
 - **Integração de dados heterogêneos:** Dados provenientes de múltiplas fontes e formatos.
 - **Qualidade dos dados:** Garantir consistência e precisão nas análises.
- **Tendências emergentes:**
 - **Data Lakes:** Integração de DWs com armazenamento de dados não estruturados.
 - **Análises em tempo real:** Processamento e consulta de dados em fluxos contínuos.
 - **Machine Learning:** Aplicações de aprendizado de máquina para previsões e automação de análises.

AWS – Amazon Redshift



Características principais:

Armazenamento em colunas:
Otimiza consultas de grandes volumes de dados.

Processamento paralelo massivo (MPP): Execução de múltiplas consultas em paralelo, aumentando a eficiência.

Escalabilidade automática:
Capacidade de adicionar ou remover nós conforme necessário.

Integração profunda com AWS:
Integração nativa com serviços como S3 (armazenamento) e Glue (ETL automatizado).



Vantagens:

Performance elevada: Ótimo para grandes volumes de dados e consultas complexas.

Custo-benefício: Boa relação para empresas já integradas ao ecossistema AWS.

Integração com Machine Learning:
Facilita a aplicação de modelos preditivos com serviços como SageMaker.



Limitações:

Complexidade de gestão: Requer conhecimentos técnicos para configuração e otimização.

Custo escalável: Pode se tornar caro para grandes volumes de dados, especialmente em ambientes com alta demanda de processamento.



Estudo de caso: McDonald's

Utiliza Redshift para análise de dados de vendas e comportamento de clientes, otimizando a personalização de ofertas e operações internas.

Microsoft Azure – Synapse Analytics



Características principais:

Integração com Power BI e Azure Machine Learning: Facilita a criação de relatórios e análises avançadas sem necessidade de mover os dados entre plataformas.

Suporte a SQL e Spark: Permite consultas em dados estruturados e não estruturados.

Escalabilidade sob demanda: Ajusta automaticamente os recursos de computação e armazenamento conforme o workload.



Vantagens:

Elasticidade: Adaptável às necessidades do negócio.

Integração com ferramentas Microsoft: Suporte nativo para Power BI, facilitando a visualização de dados.

Análises em tempo real: Processamento eficiente de grandes volumes de dados em tempo real.



Limitações:

Curva de aprendizagem: Funcionalidades avançadas, como consultas Spark, podem exigir conhecimento técnico especializado.

Custo elevado: O uso intensivo de recursos pode aumentar os custos.



Estudo de caso: Coca-Cola

A empresa usa Synapse Analytics para integrar dados de diferentes departamentos, otimizar a cadeia de abastecimento e prever a demanda de produtos.

Google Cloud Platform – BigQuery



Características principais:

Serverless: Não exige configuração manual de servidores ou clusters.

Análise em tempo real: Suporte a consultas rápidas de dados de streaming.

Machine Learning integrado: Permite o treinamento de modelos diretamente na plataforma.



Vantagens:

Facilidade de uso: Interface simples e intuitiva.

Escalabilidade automática: Ajuste dinâmico sem intervenção manual.

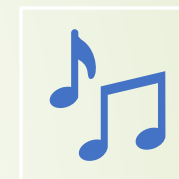
Integração com Google Analytics e Data Studio: Visualização facilitada de dados.



Limitações:

Custo de consultas: Pode se tornar caro para empresas que executam consultas frequentes em grandes volumes de dados.

Personalização limitada: Menos flexível em comparação com AWS e Azure para empresas que desejam controle total sobre a infraestrutura.



Estudo de caso: Spotify

Utiliza BigQuery para processar grandes volumes de dados de streaming e personalizar recomendações de música com base nas preferências dos utilizadores.

Comparação Geral: Custo

10



AWS Redshift:

- **Modelo de preço:** Baseado no número de nós e tempo de uso.
- **Descontos** para compromissos a longo prazo.
- **Ponto de atenção:** Custo aumenta rapidamente com a escalabilidade.

Azure Synapse Analytics:

- **Preço separado** para computação e armazenamento.
- **Flexível** para ajustar os custos conforme o workload.
- **Ponto de atenção:** Custos elevados para cargas intensivas.

Google BigQuery:

- **"Pay-as-you-go":** Paga-se por consultas e dados armazenados.
- **Ideal para uso esporádico.**
- **Ponto de atenção:** Custos podem escalar em consultas frequentes.



Comparação Geral: Escalabilidade

AWS Redshift:

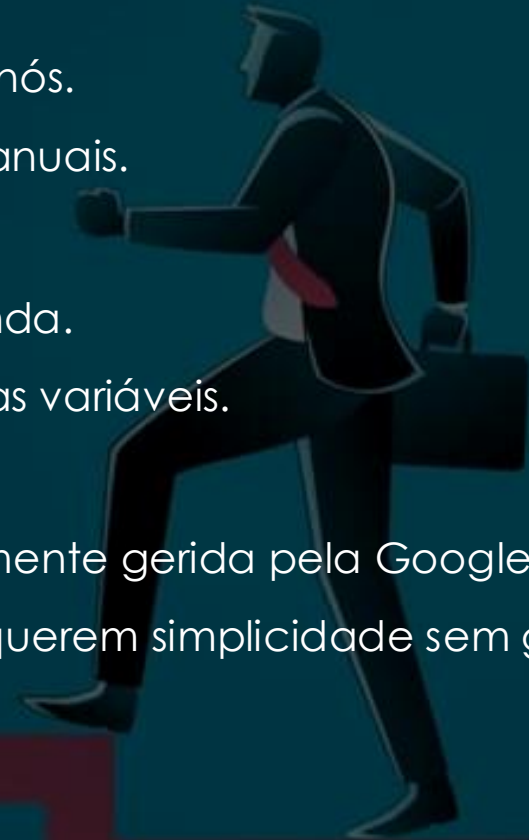
- **Escalabilidade manual:** Adição/remoção de nós.
- **Ponto de atenção:** Requer ajustes técnicos manuais.

Azure Synapse Analytics:

- **Escalabilidade automática:** Ajuste sob demanda.
- **Ponto de atenção:** Flexível, adapta-se a cargas variáveis.

Google BigQuery:

- **Escalabilidade automática (serverless):** Totalmente gerida pela Google.
- **Ponto de atenção:** Ideal para empresas que querem simplicidade sem gestão manual.



Comparação Geral: Performance

AWS Redshift:

- **Desempenho otimizado:** Armazenamento em colunas e MPP.
- **Ponto de atenção:** Desempenho depende da correcta configuração dos nós.

Azure Synapse Analytics:

- **Desempenho em tempo real:** Integração com SQL e Spark.
- **Ponto de atenção:** Requer ajuste em cargas de trabalho complexas.

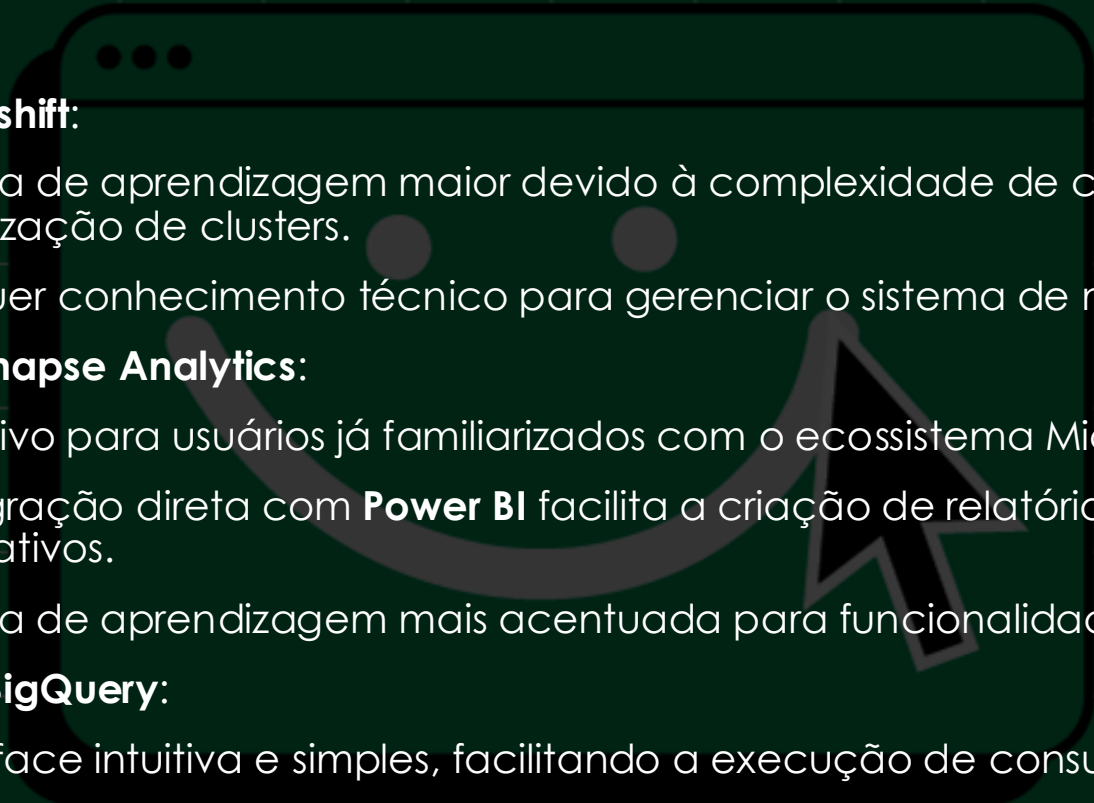
Google BigQuery:

- **Desempenho escalável:** Consultas rápidas, sem ajustes manuais.
- **Ponto de atenção:** Excelente para grandes volumes de dados, mas desempenho pode variar com a complexidade dos dados.

Comparação Geral: Segurança

- **AWS Redshift:**
 - Criptografia de dados em repouso e em trânsito.
 - Conformidade com normas como **GDPR** e **HIPAA**.
 - Integração com AWS Identity and Access Management (IAM) para controle de acesso.
- **Azure Synapse Analytics:**
 - Segurança robusta com integração ao **Azure Active Directory**.
 - Criptografia avançada de dados e conformidade com regulamentações internacionais (como **ISO/IEC 27001**).
 - Políticas de segurança ajustáveis para governança de dados.
- **Google BigQuery:**
 - Conformidade com normas de segurança como **ISO/IEC 27001**, **SOC** e **GDPR**.
 - Suporte à criptografia de dados e controle de acesso detalhado com Identity and Access Management (IAM).

Comparação Geral: Facilidade de Uso

- 
- **AWS Redshift:**
 - Curva de aprendizagem maior devido à complexidade de configuração e otimização de clusters.
 - Requer conhecimento técnico para gerenciar o sistema de maneira eficiente.
 - **Azure Synapse Analytics:**
 - Intuitivo para usuários já familiarizados com o ecossistema Microsoft.
 - Integração direta com **Power BI** facilita a criação de relatórios e dashboards interativos.
 - Curva de aprendizagem mais acentuada para funcionalidades como **Spark**.
 - **Google BigQuery:**
 - Interface intuitiva e simples, facilitando a execução de consultas SQL.
 - **Serverless:** Não requer configuração de infraestrutura, o que simplifica sua utilização.
 - Automação da escalabilidade reduz a necessidade de intervenção técnica.

Conclusão

- **Amazon Redshift:**
 - Melhor para grandes empresas que já estão no ecossistema AWS e precisam de alta performance e escalabilidade com integração profunda com outros serviços.
- **Azure Synapse Analytics:**
 - Excelente escolha para empresas que utilizam o ecossistema Microsoft, oferecendo uma solução integrada para análise e visualização de dados.
- **Google BigQuery:**
 - Ideal para empresas que desejam uma solução escalável e serverless, com foco em simplicidade e eficiência de consultas, especialmente para startups e empresas de tecnologia.
- **Conclusão final:**
 - A escolha da plataforma ideal depende das necessidades específicas de escalabilidade, custo, performance e do ecossistema tecnológico da empresa.

Referências

1. "Cloud Data Warehouse - Amazon Redshift - AWS". Amazon Web Services, Inc., aws.amazon.com/redshift.
2. "Spotify Case Study | Google Cloud". Google Cloud, cloud.google.com/customers/spotify.
3. "O McDonald's impulsiona a transformação digital na nuvem na AWS". Amazon Web Services, Inc., aws.amazon.com/pt/solutions/case-studies/mcdonalds.
4. "BigQuery enterprise data warehouse". Google Cloud, cloud.google.com/bigquery.
5. "The Coca-Cola Company and Microsoft announce five-year strategic partnership to accelerate cloud and generative AI initiatives - Stories". Stories, news.microsoft.com/2024/04/23/the-coca-cola-company-and-microsoft-announce-five-year-strategic-partnership-to-accelerate-cloud-and-generative-ai-initiatives.
6. "Cloud Data Warehouse Comparison: Redshift vs BigQuery vs Azure vs Snowflake for Real-Time Workloads". Striim, www.striim.com/blog/cloud-data-warehouse-comparison-redshift-vs-bigquery-vs-azure-vs-snowflake-for-real-time-data.
7. Smine, Sadok. "Big Data Solutions: BigQuery, Redshift, and Azure Synapse Analytics". Medium, 8 out 2023, medium.com/@sadoksmine8/big-data-solutions-bigquery-redshift-and-azure-synapse-analytics-4e842692a9f9.