Proposta de Arquitetura - Data Lake

Gustavo H. Martins

Apresentação inicial

Olá, membros do Grupo Empresarial,

Me chamo Gustavo, Engenheiro de Dados, 31 anos, Pai do Thomás, Esposo de Jéssica, Mineiro de nascença e coração, Gaucho em construção...

É com grande entusiasmo que apresento minha proposta de arquitetura para o Data Lake, um componente essencial na gestão e análise eficiente de dados para impulsionar o sucesso do Grupo Empresarial. Esta proposta visa maximizar a

eficiência no processamento, garantir escalabilidade e promover a fluidez dos dados entre diferentes camadas do pipeline.

A proposta: Construção, Sustentação e Repasse de conhecimento

1. Camada Transient: Preparando o Terreno

Na etapa Transient, os dados são armazenados temporariamente, aguardando o processamento inicial para a camada Bronze. Aqui, garantimos eficiência através da utilização de armazenamento escalável, como o Amazon S3 na AWS, em formatos verdadeiramente diversos! Este ambiente transient possibilita o processamento assíncrono e a execução de tarefas de pré-processamento, proporcionando uma base sólida para as fases subsequentes do pipeline.

2. Camada Bronze: Dados Brutos e Estruturados

Na camada Bronze, os dados são carregados de maneira bruta, porém estruturada, no formato Parquet. Isso estabelece uma base sólida para o Data Lake, possibilitando consultas eficientes e facilitando o processamento subsequente. A escalabilidade é garantida por meio da distribuição de dados em clusters, utilizando serviços como o Amazon EMR.

3. Camada Silver: Refinando a Qualidade e Tipagem de Dados

A camada Silver é o ponto em que os dados passam por tratamentos de qualidade e são tipificados de acordo com as necessidades do negócio. Utilizando ferramentas como Apache Spark e AWS Glue, asseguramos a eficiência no processamento, realizando limpeza, enriquecimento e transformação. A escalabilidade é mantida por meio da automação de tarefas e da adaptação dinâmica a variações de carga.

4. Camada Gold: `Transformando Dados em Insights de Negócios`

Na camada Gold, os dados são cruzados com fontes provenientes da camada Silver para responder a perguntas de negócios. Utilizando técnicas avançadas de processamento analítico, como consultas SQL otimizadas e machine learning, asseguramos que o Data Lake se torne uma fonte valiosa de insights estratégicos. A escalabilidade é mantida por meio de arquiteturas de data warehousing eficientes, como o Amazon Redshift.

5. Conclusão: `Alcançando o Sucesso com Dados Confiáveis e Estratégicos`

Em síntese, a proposta de arquitetura para o Data Lake no Grupo Empresarial tem como objetivo estabelecer um ambiente eficiente, escalável e dinâmico. A adoção das camadas Transient, Bronze, Silver e Gold reflete nosso compromisso em transformar dados em ativos estratégicos, fornecendo uma infraestrutura sólida para tomadas de decisão informadas e bem-sucedidas. Este modelo busca não apenas gerenciar, mas potencializar a riqueza dos dados, impulsionando o sucesso e a inovação no âmbito do Grupo Empresarial.

6. Observações finais:

A solução proposta, tendo a AWS como exemplo de provedor de serviços em nuvem, oferece a flexibilidade de integração com o AWS Athena. Para viabilizar essa integração, é necessário realizar o mapeamento e a definição dos catálogos de dados em cada Delta Table. Essa abordagem permite uma transição suave e eficiente entre as camadas de processamento, garantindo a interoperabilidade e maximizando a utilidade do AWS Athena no contexto do `Data Lake.`

SBS: `Sempre Bom Saber`

- Stacks utilizadas:

1. # Apache Spark v3.4.0:

O `Apache Spark` é um mecanismo `analítico` unificado para processamento de dados em grande escala. Ele fornece APIs de alto nível em Java, Scala, Python e R e um mecanismo otimizado que oferece suporte a gráficos de execução geral. Ele também oferece suporte a um rico conjunto de ferramentas de nível superior, incluindo `Spark SQL` para SQL e processamento de dados estruturados.

2. # Delta Lake:

Delta Lake é uma estrutura de armazenamento de código aberto que permite construir uma arquitetura Lakehouse com mecanismos de computação, incluindo Spark, PrestoDB, Flink, Trino e Hive e APIs para Scala, Java, Rust e Python.

3. # Java Rumtime:

Oracle Java é a linguagem de programação e plataforma de desenvolvimento nº 1. Reduz custos, encurta os prazos de desenvolvimento, impulsiona a inovação e melhora os serviços de aplicativos. Com milhões de desenvolvedores executando mais de 60 bilhões de Máquinas Virtuais Java em todo o mundo, Java continua a ser a plataforma de desenvolvimento preferida de empresas e desenvolvedores.

4. # Python:

Python é uma linguagem de programação que permite trabalhar rapidamente e integrar sistemas de forma mais eficaz.

5. # Pandas:

Em 2008, o desenvolvimento do pandas começou na AQR Capital Management. No final de 2009, ele era de código aberto e hoje é apoiado ativamente por uma comunidade de indivíduos com ideias semelhantes em todo o mundo, que contribuem com seu valioso tempo e energia para ajudar a tornar possíveis os pandas de código aberto. Obrigado a todos os nossos colaboradores.

Desde 2015, o pandas é um projeto patrocinado pela NumFOCUS . Isto ajudará a garantir o sucesso do desenvolvimento do pandas como um projeto de código aberto de classe mundial.

6. ## DuckDB:

DuckDB é um sistema de gerenciamento de banco de dados SQL OLAP em processo. clique aqui e verá porque eu (Gustavo) acredito que o DuckDB vai mudar muita coisa num futuro breve.

Agradecimentos:

Estou empolgado com a perspectiva de colaborar para o progresso tecnológico do Grupo Empresarial por meio desta parceria. Estou à disposição para aprofundar as discussões e esclarecer quaisquer dúvidas que possam surgir. Atenciosamente,

Gustavo H. Lopes Contato

Autor:

- Gustavo H Martins (GitHub | LinkedIn)