Proposta de Arquitetura - Data Lake Grupo Panvel

Gustavo H. Martins

Apresentação inicial

Olá, membros do Grupo Panvel,

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 Panvel. 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 camada Transient, os dados são armazenados temporariamente, aguardando o processamento inicial para a camada Bronze. Aqui, a eficiência é garantida por meio da utilização de armazenamento escalável, como o Amazon S3 na AWS, em formatos diversos, diversos mesmo! Esse ambiente transient possibilita o processamento assíncrono e a execução de tarefas pré-processamento.

2. Camada Bronze: `Dados Brutos e Estruturados`

Na camada Bronze, os dados são carregados de forma bruta, porém estruturada, no formato `Parquet`. Isso proporciona uma base sólida para o Data Lake, permitindo consultas eficientes e facilitando o processamento subsequente. A escalabilidade é assegurada 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 onde os dados recebem tratamentos de qualidade e são tipados de acordo com as necessidades do negócio. Utilizando ferramentas como Apache Spark e AWS Glue, garantimos a eficiência no processamento, realizando limpeza, enriquecimento e transformação. A escalabilidade é mantida através 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, garantimos 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 resumo, esta proposta de arquitetura para o Data Lake no Grupo Panvel visa criar um ambiente eficiente, escalável e fluido. Ao adotar a abordagem de camadas Transient, Bronze, Silver e Gold, garantindo que os dados se transformem em ativos estratégicos para tomadas de decisão informadas e bem-sucedidas.

6. Observações finais:

A solução apresentada, foi usado como exemplo a provedora de cloud AWS, neste contexto, cada camada de processamento pode ser também consumida pelo `AWS Athena`, basta que para isso façamos o mapeamento e definição dos catálogos de dados a serem consumidos em cada `Delta Table.`

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 entusiasmado com a possibilidade de contribuir para o avanço tecnológico do Grupo Panvel por meio desta parceria.

Fico à disposição para discussões adicionais e esclarecimento de dúvidas.
Atenciosamente,

Gustavo H. Lopes Contato

Autor:

- Gustavo H Martins (GitHub | LinkedIn)