**1. Considerando que a Ame possui diferentes aplicações, resultando em diferentes fontes**

**de dados (como bancos relacionais e noSQL), de que maneira você construiria uma**

**arquitetura para realizar a ingestão desses dados em uma plataforma de Big Data?**

**Descreva as tecnologias que você escolheria para realizar a ingestão, bem como o fluxo**

**de dados entre elas (lembrando que o objetivo é disponibilizar as informações o mais**

**próximo de "real-time" possível).**

**R:** Atualmente temos bastante tecnologias que nos permite realizar a ingestão de dados de várias fontes, no entanto alguns outros fatores são importantes a se levar em conta no desenho da arquitetura (além da parte tecnológica). Considerando que o time de Devops da Ame digital seja maduro e suficiente para suportar uma arquitetura open source, eu recomendaria a seguinte arquitetura:

Diagrama

Descrição gerada automaticamente

Principais motivos para a escolha:

* Atualmente o Kafka connect tem suporte para vários datasources (sql e nosql).
* Kafka está amplamente conhecido na comunidade, facilitando o onboarding entre os colaboradores da equipe e novos colaboradores.
* Existe um operator no kubernetes chamado Strimzi que facilitaria a gestão de uma arquitetura cross-cloud com Kafka.
* O ksqlDB é super compatível com Kafka, atua com transformações “like sql” e opera em streaming (requisito da arquitetura).
* Com o dado transformado de volta para o kafka, pode-se compartilhar facilmente outros consumidores (micro serviços por exemplo).

Para uma arquitetura mais cloud (Saas, Paas) teríamos alguns fatores como a cloud predominante na Ame digital, geralmente o executivo de contas de uma Cloud consegue negociar valores para os produtos. Desta forma considerando uma primeira arquitetura de ingestão de dados em tempo real com uma alta disponibilidade eu escolheria a plataforma da Confluent pois ela opera de forma cross-cloud dentro dos marketplaces da Azure, AWS e GCP o que facilitaria uma replica em outras regiões ou outros providers. Em segundo plano eu optaria por uma arquitetura AWS pela ampla participação de mercado e do comunicado em cenários de bigdata, utilizando a seguinte stack:

Diagrama

Descrição gerada automaticamente

**2. Ao utilizar ferramentas de processamento distribuído como Spark ou Hive, é muito**

**comum enfrentar problemas relacionados à má distribuição de dados entre as máquinas**

**do cluster, diminuindo drasticamente a performance das aplicações, principalmente em**

**operações relacionadas a agregação ou join. Utilizando seus conhecimentos e**

**experiências, descreva uma possível solução para o problema em questão.**

**R:** Com Spark não se deveria ter problemas de performance, principalmente de uma tecnologia de computação distribuída, deve se analisar como esta o cluster spark e como está sendo desenvolvimento as analises e processamento dos dados (algoritmo).

Algum dos fatores que contribui principalmente para a má distribuição de dados entre as máquinas (workers) do cluster é a forma como os dados são processados, embaralhados (shuffle) e atribuídos tarefas de cada partição ( isso pode ser observado no spark UI no stage de processamento com o problema) ou seja, ajustando o número de partições (repartition) pode-se resolver o problema visto que cada partição poderá ter uma quantidade de dados mais equilibrada para as tarefas de processamento. Outra alternativa é analisar se existe conjunto de dados menores com junções para dados maiores ( Data broadcast ), alterando um parâmetro de broadcastjoin pode-se resolver o problema da má distribuição. Por último as vezes necessita de mais pré-processamento, principalmente se existe uma operação em sequencia com muitas transformações para dados desnecessários que não são do interesse final (exemplo valores nulos).

**3. O dia a dia de um engenheiro de dados, dentre outras tarefas, é disponibilizar as**

**informações em alta performance (próximos a real-time) para Analistas e Cientistas de**

**Dados de modo a possibilitar à análise e criação de modelos estatísticos. De que modo**

**e quais tecnologias você usaria para disponibilizar os dados para estas pessoas.**

**R:**

Sem sombra de dúvidas através de algum query engine (Athena, bigquery, dremio) sobre um Data Lake, primeiro ponto é a performance que se pode ter com alguma dessas tecnologias, segundo seria para fins de governança e catalogo de dados (visto que estaria tudo no datalake) e por ultimo a facilidade de coletar os dados com linguagem SQL.

**4. Por fim, tendo em mente o crescimento exponencial dos dados e utilização massiva da**

**plataforma de Big Data, quais métodos de organização e/ou governança você**

**implementaria para manter o ambiente sustentável?**

**R:**

Implementaria um processo de auditoria de implantação de novas pipelines para atribuir alguns requisitos de implantação no que tange a manipulação do datalake, um dos requisitos pode ser a obrigatoriedade de atualizar o catalogo de dados, avaliar a estrutura de pasta e camadas planejadas para aquela ingestão/transformação e por ultimo escolheria uma ferramenta o mais abrangente nesse universo de governança/catalogo de dados, tenho conhecimento do Apache atlas para governança e catalogo de dados.