Solução

O fluxo de dados nesta arquitetura começa com a captura dos eventos gerados pelos dispositivos dos usuários, passando por etapas de ingestão, processamento, e armazenamento, até chegar na disponibilização para consumo de outras áreas.

1. Fonte de Dados: Streaming e Batch

- Streaming: Os eventos gerados pelos dispositivos são enviados em tempo real sempre que ocorre uma ação do usuário, como assistir um vídeo, pausar, ou adicionar aos favoritos. Os eventos são enviados para o Google Cloud Pub/Sub, que atua como um Message Broker publicando os eventos em determinados tópicos para serem processados. A alta taxa de ingestão é suportada pelo balanceamento de carga embutido no Pub/Sub.
- Batch: Também há a possibilidade de eventos serem enviados de maneira agendada, como logs e dados agregados que são coletados em lote. Esses dados são armazenados no Google Cloud Storage em intervalos programados para serem processados.

2. Ingestão de Dados: Pub/Sub e Cloud Storage

- Pub/Sub: Os eventos em tempo real enviados para o Pub/Sub são assinados por diferentes serviços para processamento. Se houver falhas durante a ingestão, políticas de Retry são aplicadas para garantir que os eventos sejam reprocessados. Circuit Breakers estão implementados para evitar sobrecargas e falhas em cascata, isolando serviços que estejam temporariamente indisponíveis.
- Cloud Storage: Os dados em batch armazenados no Cloud Storage são processados e se necessario acionados manualmente ou por agendamento usando Cloud Composer. Suportando diferentes formatos, como JSON, Parquet ou CSV. Ele atua como o ponto de entrada para processamento em lote.

3. Processamento: Dataflow e Dataproc

- Dataflow: Para dados em streaming, o Dataflow processa os eventos em tempo real. Ele realiza transformações, como agregação de eventos, limpeza de dados, ou enriquecimento de informações, antes de armazenar os dados processados no BigQuery ou Cloud Bigtable. A observabilidade é garantida por Stackdriver (Google Cloud Operations), que monitora o desempenho e a saúde dos pipelines. Em caso de falhas, são aplicadas políticas de Retry e Circuit Breaker para assegurar a continuidade do processamento.
- Dataproc: Dados em batch, como aqueles armazenados no Cloud Storage, são processados em grande escala usando Dataproc. O Dataproc lida com tarefas complexas de big data, como transformações massivas ou cálculos distribuídos. Stackdriver também é usado para monitorar e alertar sobre o status dos jobs do Dataproc.

4. Data Warehouse: BigQuery

- BigQuery: BigQuery é o data warehouse centralizado, otimizado para consultas rápidas e eficientes em grandes volumes de dados. Ele permite o armazenamento e consulta de dados em tempo real, além de integração com outras ferramentas do Google Cloud. O uso de partições e clustering permite que consultas sejam otimizadas para desempenho. Para consultas frequentes, Google Cloud Memorystore (Redis) é utilizado como cache para armazenar resultados, reduzindo a latência e aliviando a carga.

5. Databases: Cloud Bigtable

- Cloud Bigtable: Dados que necessitam de baixa latência e alta taxa de transferência, como logs de eventos ou dados de sessão de usuários, são armazenados no Cloud Bigtable. O Memorystore também pode ser usado para cachear dados acessados frequentemente a partir do Bigtable, melhorando ainda mais a performance.
- Cloud SQL: Gerenciar dados relacionais estruturados. É ideal para armazenar informações transacionais e dados que necessitam de consistência

6. Machine Learning: Vertex AI

- Vertex AI: Vertex AI é a plataforma gerenciada para construir, treinar, e implantar modelos de machine learning. Através de datasets processados, a área de ciência de dados pode criar modelos de recomendação ou análise de comportamento de usuários.

7. Visualização: Looker, Data Studio e GKE

- Looker e Data Studio: Visualização dos dados, criando dashboards e relatórios para a área de negócios. Looker oferece uma plataforma mais robusta com modelagem de dados integrada, enquanto o Data Studio permite relatórios de fácil criação.
- GKE (Google Kubernetes Engine): Para casos mais complexos, o GKE pode ser usado para hospedar aplicações customizadas de visualização ou análise, permitindo escalabilidade e flexibilidade.

8. Orquestração: Workflows, Cloud Composer, Data Catalog e Cloud Run

- Workflows e Cloud Composer: Essas ferramentas orquestram o fluxo de dados, gerenciando pipelines, rotinas de processamento em batch, e automação de tarefas.
- Data Catalog: O Data Catalog ajuda a gerenciar metadados, facilitando a descoberta e a governança dos dados. Ele é crucial para a classificação e gestão de dados sensíveis, garantindo que o acesso aos dados seja restrito e controlado.
- Cloud Run: Usado para a execução de serviços baseados em contêineres, especialmente para APIs ou microserviços que precisam ser acionados de maneira flexível e escalável.

9. Governança e segurança:

- IAM: Gerenciar o acesso e as permissões de usuários e serviços ao longo de toda a arquitetura, como por exemplo acesso ao BigQuery, Cloud Storage, Data Catalog, etc, garantindo que apenas usuários e serviços autorizados possam acessar ou modificar os dados.
- Data Loss Prevention: Detectar e proteger dados sensíveis, como informações de identificação pessoal (PII), antes que sejam armazenados ou processados, sua utilização na camada de ingestão e processamento será para identificar e mascarar os dados sensíveis.