Análise das Propostas de Infraestrutura de Cloud para Análise de Dados

1. Contextualização

A Prefeitura está passando por um processo de transformação digital em que o uso de dados para definição de políticas públicas é essencial. Considerando não haver uma referência técnica definida e instalada nessa área, faz-se necessário definir a Arquitetura em Nuvem para Projetos de Dados, de modo que se possa dispor de uma infraestrutura para processar e organizar esses dados que consiga escalar e se adaptar às demandas da Prefeitura.

Para isso, o Escritório de Dados - GBP e a Subsecretaria de Governo Digital e Transparência - SEGOVI estudaram soluções encontradas no mercado e em outros entes do governo para definir a arquitetura necessária. A arquitetura é estruturada em componentes que interagem entre si e completam o ciclo de dados desde a extração em seu formato bruto até o uso em soluções de BI. Cada componente, no que lhe concerne, tem requisitos específicos para atender às demandas dos órgãos e entidades da Prefeitura do Rio de Janeiro. A apresentação completa desses componentes e seus requisitos está no Anexo 1 deste documento.

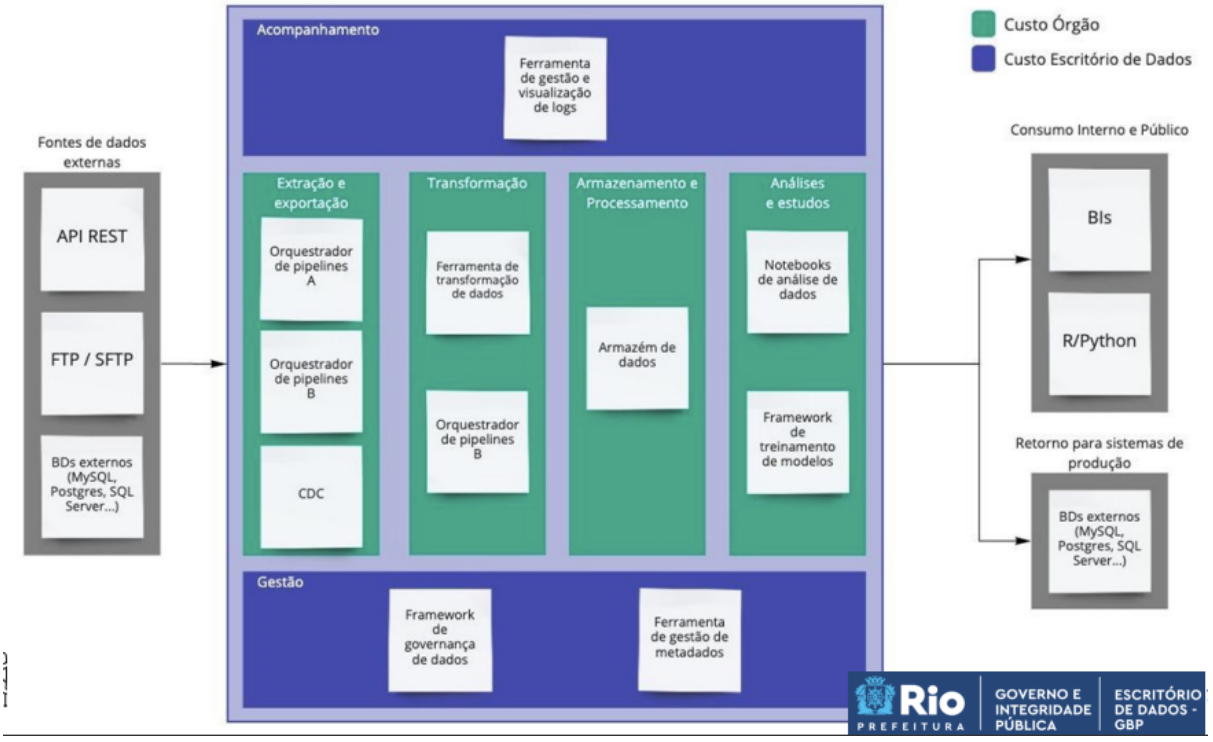


Figura 1. Diagrama com componentes da infraestrutura proposta. Cada componente tem uma lista de requisitos

Esse estudo foi feito por convocação via consulta pública, com ampla divulgação, tanto em mídia impressa, quanto no portal da prefeitura. A mesma ocorreu no dia 16/12/2021 com a participação de 7 empresas. O objetivo nessa data foi apresentar a arquitetura pretendida para que os fabricantes pudessem propor seus componentes, que atenderiam às demandas, com respectivas estimativas de custos.

A apresentação realizada na consulta pública está disponível nesse sítio: https://segovi.prefeitura.rio/transparencia/. Na apresentação, estava disponível um link para uma planilha contendo os itens da arquitetura que cada fabricante deveria atender, a qual deveria ser preenchida com seus respectivos componentes e encaminhada em retorno até dia 10/01/2022, incluindo estimativa de preços.

A Figura 2 sumariza as respostas das cinco empresas que responderam no prazo estipulado: Extreme Digital (Google Cloud), Informatica (Azure), Amazon Web Services, VSDATA (IBM) e TIBCO.

2. Comparação

		OPEX		Componentes	Requisitos		
	Importância do Componente	Mensal	Anual	Total	Completos	Parciais	Não Cumpridos/Não Especificados
Google Cloud	Obrigatório	R$26,447.75	R$317,373.00	7	100.00%	0.00%	0.00%
	Todos	R$57,065.95	R$684,791.40	8	100.00%	0.00%	0.00%
Azure	Obrigatório	-	-	5	93.22%	6.78%	0.00%
	Todos	R$121,340.54*	R$1,456,086.48*	7	93.42%	6.58%	0.00%
Amazon Web Services	Obrigatório	R$25,026.84*	R$300,322.08*	7	89.83%	10.17%	0.00%
	Todos	R$53,582.69*	R$642,992.26*	10	92.11%	7.89%	0.00%
IBM Cloud	Obrigatório	Não Informado	Não Informado	Não está claro	89.83%	6.78%	3.39%
	Todos	Não Informado	Não Informado	Não está claro	90.79%	5.26%	3.95%
TIBCO	Obrigatório	-	-	5	98.31%	1.69%	0.00%
	Todos	R$1,212,726.21	R$14,552,714.58	7	92.11%	3.95%	3.95%

Figura 2. Tabela comparando as respostas dos requerentes ordenados pelo cumprimento dos requisitos. (*) Custos estimados pela equipe do Escritório de Dados usando as calculadoras disponíveis fornecidas pelos provedores.

Os critérios de avaliação foram divididos por cumprimento dos requisitos técnicos e por custo. Para cada critério a solução ganhara um *score* (máximo 5, mínimo 1) referente a posição dela em relação às outras soluções (Figura 3). Os requisitos são os seguintes:

1) Maior número de requisitos obrigatórios completos

A plataforma deve obrigatoriamente prover esses requisitos. Caso contrário, entendemos que novas tecnologias deverão ser incluídas para satisfazer a arquitetura desejada, incorrendo em maior custo total da solução.

2) Maior número de requisitos totais completos

A plataforma que satisfizer mais requisitos melhor se adequa às preferências da arquitetura proposta. Isso permite que a equipe da prefeitura possa desenvolver exclusivamente em um só ambiente, diminuindo o custo de capacitação e transferência de conhecimento. Caso necessário, esse será levado como critério de desempate.

3) Menor número total de componentes

Uma quantidade reduzida de componentes significa um menor custo de manutenção da infraestrutura. A equipe técnica da prefeitura é diminuta e com capacidade limitada para oferecer manutenção. Portanto, uma solução que atende os requisitos obrigatórios, mas com um número menor de componentes será preferida.

4) Menor preço

Uma vez com os requisitos técnicos analisados, a solução com menor preço será considerada.

Empresa	Maior número de requisitos obrigatórios	Maior número de requisitos totais	Menor número total de componentes	Soma
Google Cloud	5	5	4	14
Azure	3	4	5	12
TIBCO	4	3	5	12
Amazon Web Services	2	3	3	8
IBM Cloud	2	2	2	6

Figura 3. Tabela sumarizando os *scores* de cada provedor.

Resultados

1) Maior número de requisitos obrigatórios completos

A Figura 2 está ordenada por número de requisitos obrigatórios preenchidos. A única empresa que preenche todos os requisitos é a Google Cloud. Todas as outras preenchem satisfatoriamente a lista de requisitos obrigatórios, sendo a IBM e AWS as com menor preenchimento de 89.83%.

2) Maior número de requisitos totais completos

A Google Cloud também é a única que preenche todos os requisitos necessários. Neste caso, as soluções da Azure, AWS e TIBCO preenchem satisfatoriamente com mais de 90%.

3) Menor número total de componentes

Tanto a TIBCO quanto a Azure são os provedores com menor número de componentes, seja no total de componentes ou nos obrigatórios. A Google Cloud tem 1 componente a mais no total de componentes, já a AWS chega a ter 10 componentes na solução completa.

4) Menor preço

A AWS é a solução que ofereceu menor preço, com uma diferença de R$1.441 no valor mensal proposto pela Extreme Digital (Google Cloud). Informatica (Azure) e TIBCO apresentaram valores bem acima dos concorrentes. Já a IBM não apresentou uma proposta de precificação dos componentes. É importante ressaltar que Informatica (Azure) e AWS não forneceram estimativas de valor, porém as soluções dispõem de uma ferramenta online que permite fazê-las. Esse processo de estimação do valor foi feita pela equipe do Escritório de Dados e não por representantes das soluções.

Conclusão

A Google Cloud foi a plataforma que teve a maior pontuação na classificação da Figura 3. Ela oferece todos os requisitos obrigatórios e opcionais, um número reduzido de componentes e um preço competitivo. Entendemos que soluções que não ofereceram todos os requisitos terão que ser complementadas com outras ferramentas. Esse acréscimo aumentará o número de componentes e, consequentemente, o preço. Além de ter um custo técnico implícito de adequar a equipe a outro ambiente de desenvolvimento. A AWS é a solução com maior economicidade, porém, por não oferecer todos os requisitos, outros componentes terão que ser adicionados. Como a diferença de preço entre a AWS e Google Cloud é baixa, entendemos que não existe uma economia real em optar pela AWS.

Portanto, a Google Cloud será a solução oferecida para os órgãos da prefeitura no escopo de projetos de dados e também a plataforma oficial de desenvolvimento de infraestrutura de dados do Escritório de Dados.

3. Cumprimento de requisitos técnicos

1.1. Captura de dados alterados (CDC)

As soluções propostas por todos os participantes, cada qual com suas particularidades, parecem suprir as necessidades do Escritório Municipal de Dados.

1.2. Orquestrador de Pipelines A

A solução proposta utilizando a ferramenta TIBCO Data Migrator não cumpre o primeiro requisito apresentado e, portanto, torna-se inviável.

A capacidade de integração direta com Discord aparentemente foi um requisito que alguns participantes, como AWS, Informatica e IBM, marcaram como parcialmente cumprido ou não cumprido. Como é possível implementar tal funcionalidade em Python, dado que o suporte a Python deve ser pleno, o descumprimento desse requisito não torna a ferramenta inviável.

As outras soluções não citadas nessa seção suprem as demandas apresentadas.

1.3. Orquestrador de Pipelines B

As soluções propostas por todos os participantes, cada qual com suas particularidades, parecem suprir as necessidades do Escritório Municipal de Dados.

1.4. Ferramenta de transformação de dados

Os participantes AWS e Informatica descrevem o primeiro requisito como um desenvolvimento à parte. No entanto, era esperado que essa característica estivesse compreendida no conjunto de ferramentas propostas.

Além disso, com relação à AWS, o segundo requisito foi marcado como parcialmente atendido sem mais explicações.

A participante IBM também registrou o primeiro requisito como parcialmente atendido sem explicações

As outras soluções não citadas nessa seção suprem as demandas apresentadas.

1.5. Armazém de dados

A solução proposta apresentada pela TIBCO menciona o ecossistema Hadoop, porém não explicita como essas ferramentas serão hospedadas na nuvem e como interagir com as mesmas.

O apresentado pela Informatica não possui integração com o Apache Superset, além de ser necessário recorrer a uma plataforma não focada em *big data* para realização de análises geoespaciais.

A IBM atende parcialmente os requisitos 2, 8 e 11, sem maiores explicações.

As outras soluções não citadas nessa seção suprem as demandas apresentadas.

1.6. Ferramenta de gestão e visualização de logs

As soluções propostas por todos os participantes, cada qual com suas particularidades, parecem suprir as necessidades do Escritório Municipal de Dados.

1.7. Notebooks de análise de dados

A Extreme Digital, apesar de ter especificado níveis diferentes na plataforma apresentada para os kernel tipo A e B, não especificou dimensionamento para o kernel tipo C.

Além disso, a TIBCO não cumpre os requisitos 1, 3 e 5, além de cumprir somente parcialmente os requisitos 2 e 7.

De forma similar, a IBM não cumpre o requisito 1.

As outras soluções não citadas nessa seção, ainda que apresentem algum requisito parcialmente cumprido, foram consideradas suficientes às demandas apresentadas.

1.8. Framework de treinamento de modelos

As soluções propostas por todos os participantes, cada qual com suas particularidades, parecem suprir as necessidades do Escritório Municipal de Dados.

1.9. Framework de governança de dados

As soluções propostas por todos os participantes, cada qual com suas particularidades, parecem suprir as necessidades do Escritório Municipal de Dados.

1.10. Ferramenta de gestão de metadados

A IBM não apresenta suporte à geração automática de documentação através dos metadados.

As outras soluções não citadas nessa seção, ainda que apresentem algum requisito parcialmente cumprido, foram consideradas suficientes às demandas apresentadas.

gov.br Documento assinado digitalmente
JOAO LUIZ MARTINS CARABETTA
Data: 13/04/2022 10:39:36-0300
Verifique em https://verificador.iti.br

João Luiz Martins Carabetta
Chefe Executivo de Dados

Anexo 1 - Apresentação descrevendo requisitos

Agenda

- Objetivo
- Apresentação do Data Lake
  - **O que** queremos resolver
  - **Como** estamos resolvendo
  - **Com o que** estamos resolvendo
- Aprofundamento
  - **Coordenação** de componentes
  - **Casos** de uso
  - **Próximos** passos
  - **Perguntas**

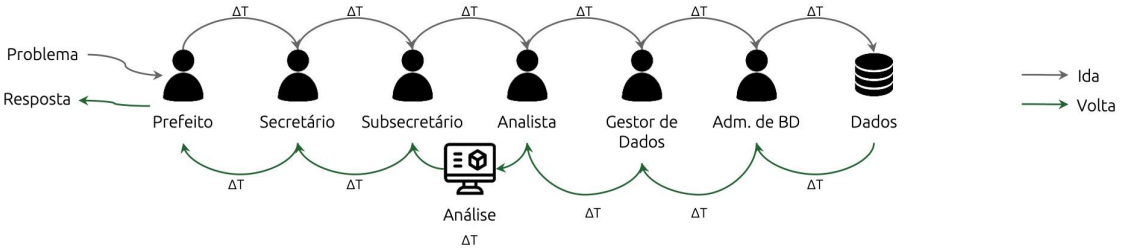
Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP

Objetivo

- A presente Consulta Pública tem como propósito a definição técnica de uma arquitetura em nuvem para projetos de dados.
- Este documento, composto da descrição de uma arquitetura tecnologicamente agnóstica e de um anexo de requisitos, apresenta a visão do problema que se espera resolver e para o qual interessados podem apresentar suas soluções para apreciação interna.
- As propostas de soluções deverão ser preenchidas a partir de uma cópia dessa planilha.
- A partir das informações apresentadas a Prefeitura do Rio realizará uma análise técnica para definição dos elementos que sustentarão a arquitetura estratégia para projetos de dados.

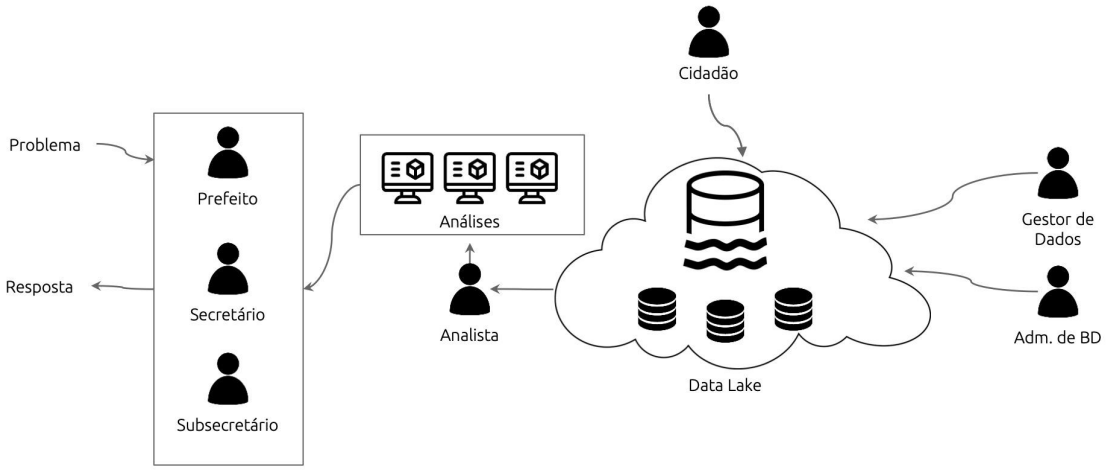
O que queremos resolver

- Demora para acessar dados
- Dificuldade em localizar dados
- Dados inconsistentes
- Tecnologias desatadas: difícil de trabalhar com diferentes conjuntos de dados
- Baixa escalabilidade: escala verticalmente



Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP

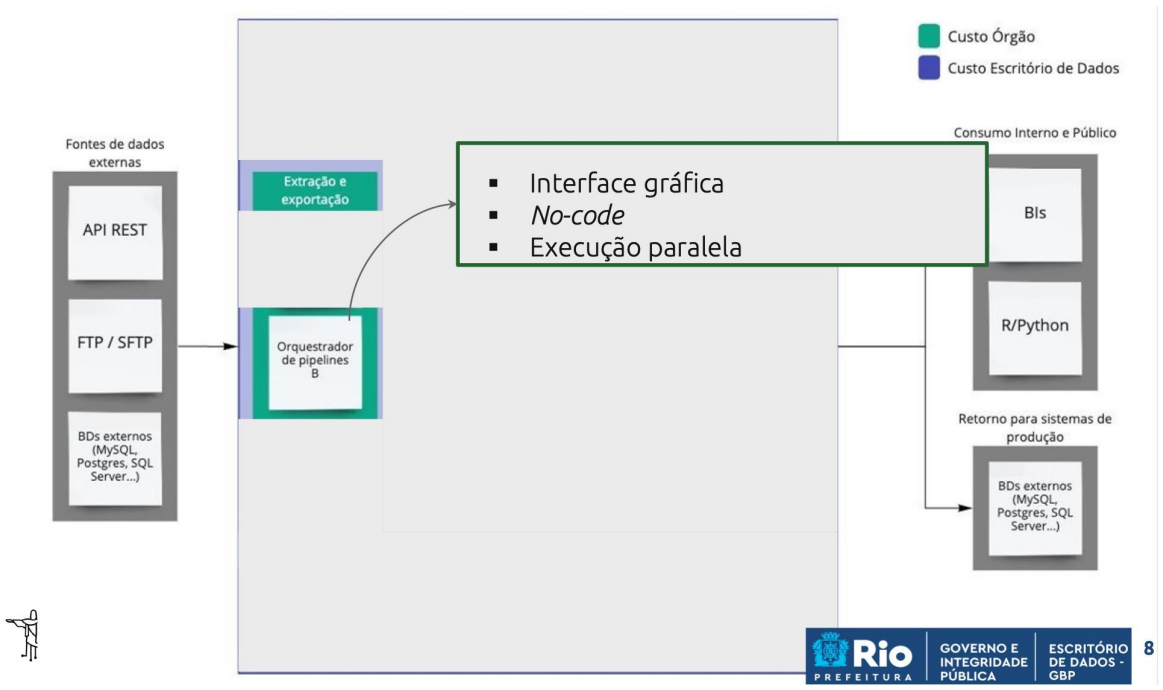
Como estamos resolvendo

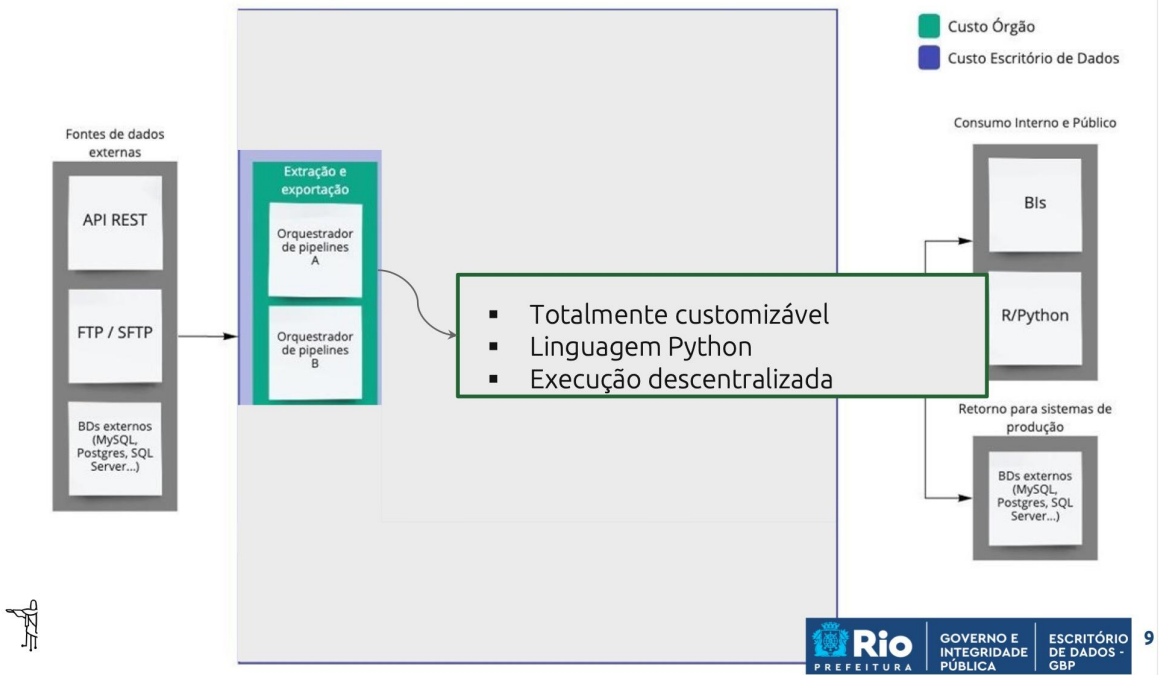


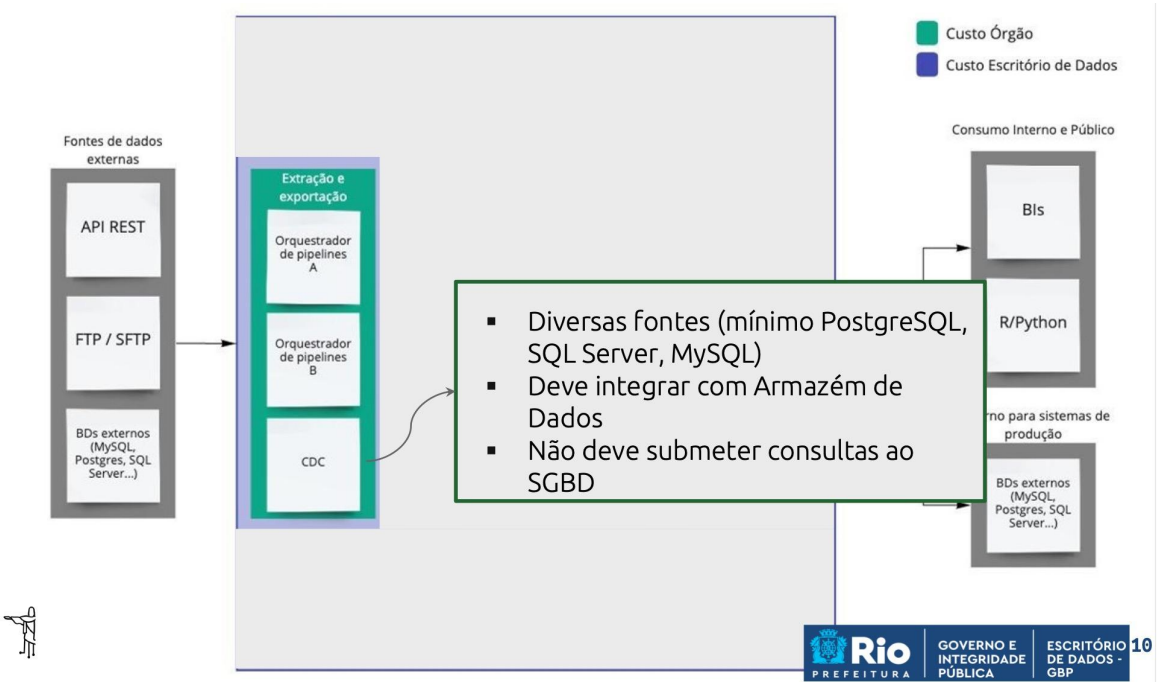
Como estamos resolvendo

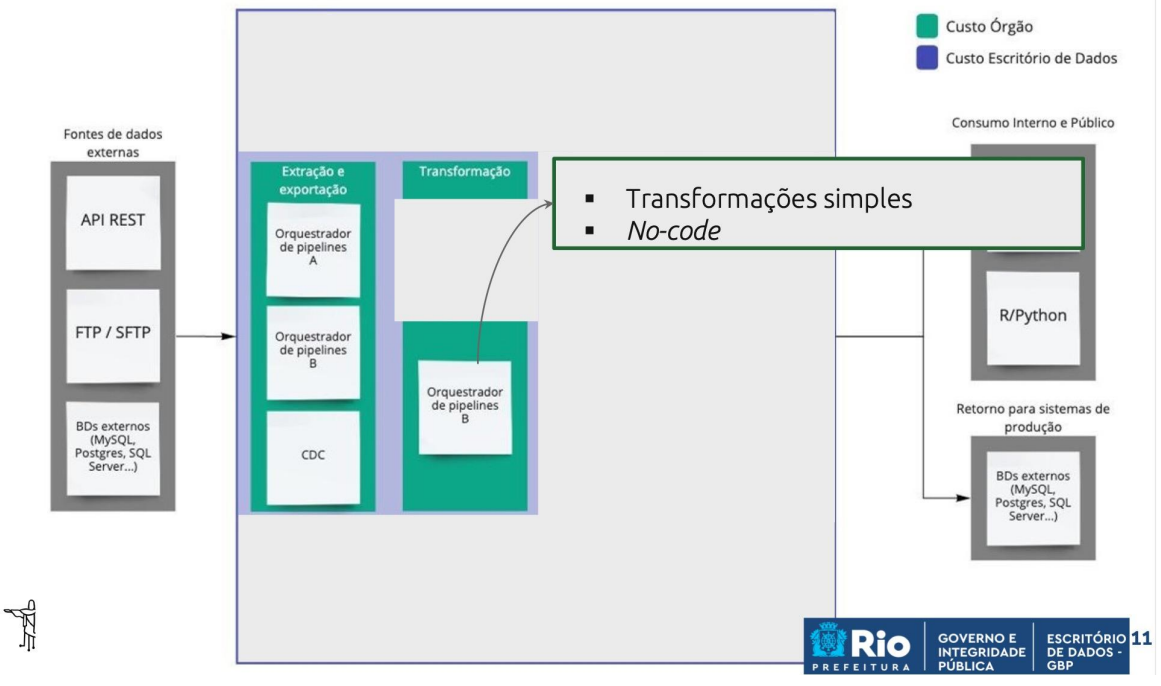
Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP

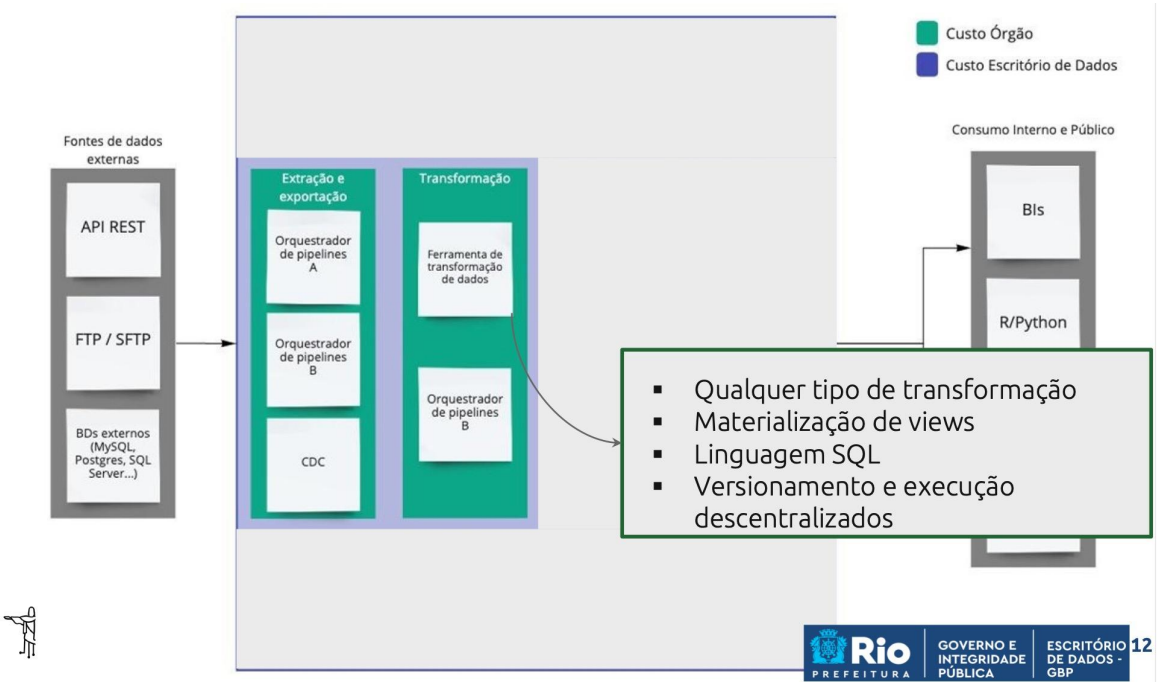
- **Custo descentralizado** e dimensionado em função do uso
- **Dados centralizados** e com uma única interface de acesso
- **Padronização, harmonização e qualidade dos dados**
- **Unidade de tecnologias** utilizadas para extração e processamento de dados
- **Padronização de estilos de código**
- **Gerenciamento de acesso aos dados** com possibilidade de disponibilizar publicamente
- **Capacidade de desacoplamento** da macroestrutura, caso necessário
- **Infraestrutura modular, escalável e de fácil manutenção**

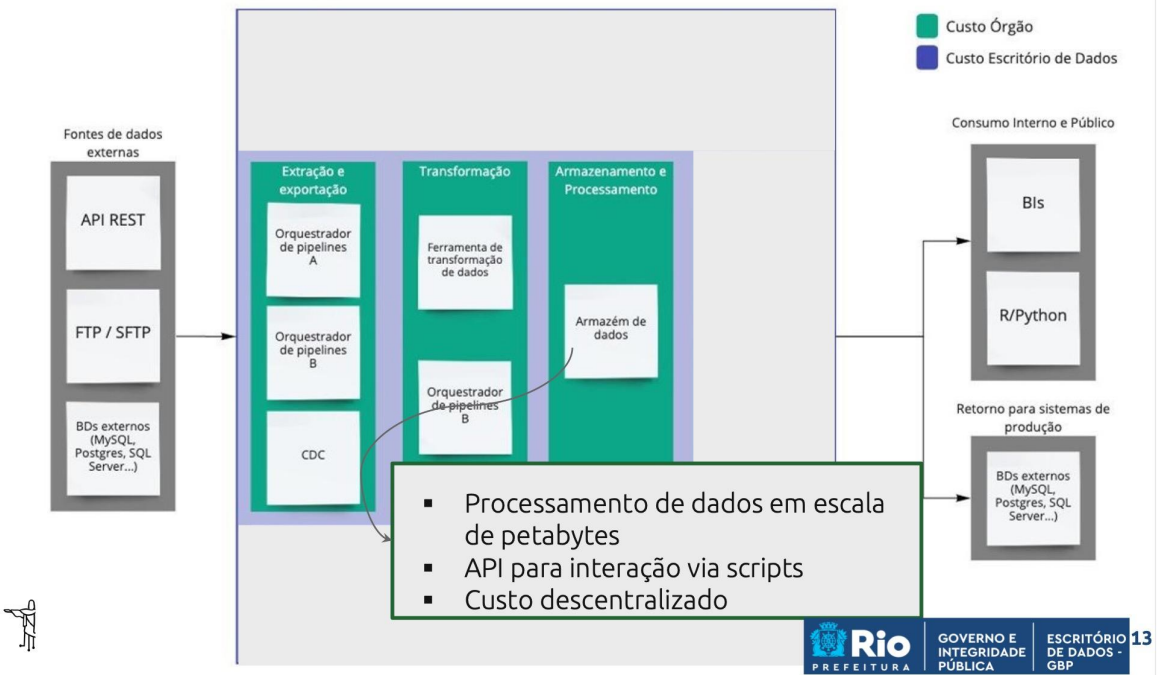

Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines B
Interface gráfica
No-code
Execução paralela
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)

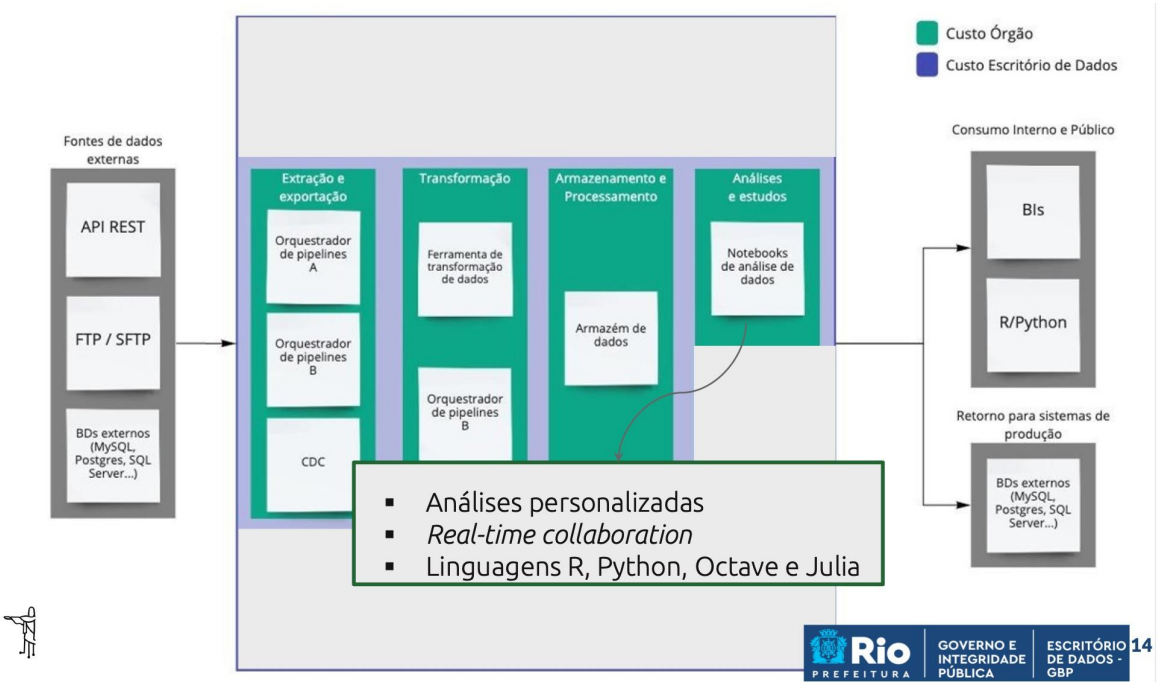

Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines A
Orquestrador de pipelines B
Totalmente customizável
Linguagem Python
Execução descentralizada
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)

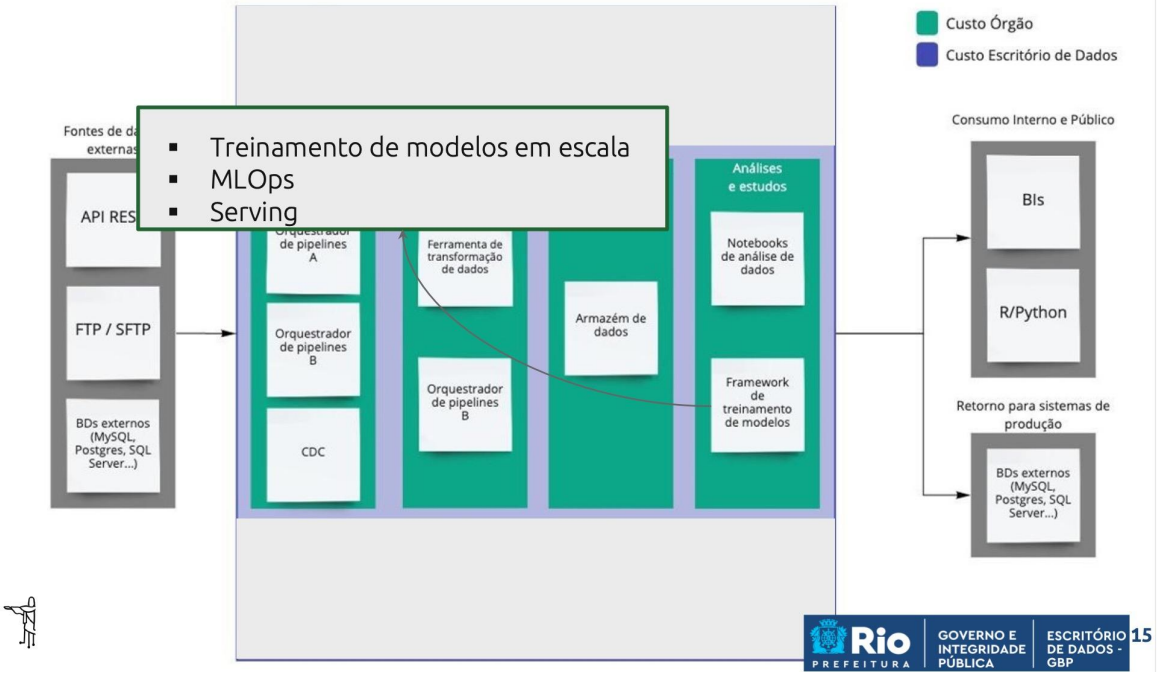

Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines A
Orquestrador de pipelines B
CDC
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Diversas fontes (mínimo PostgreSQL, SQL Server, MySQL)
Deve integrar com Armazém de Dados
Não deve submeter consultas ao SGBD

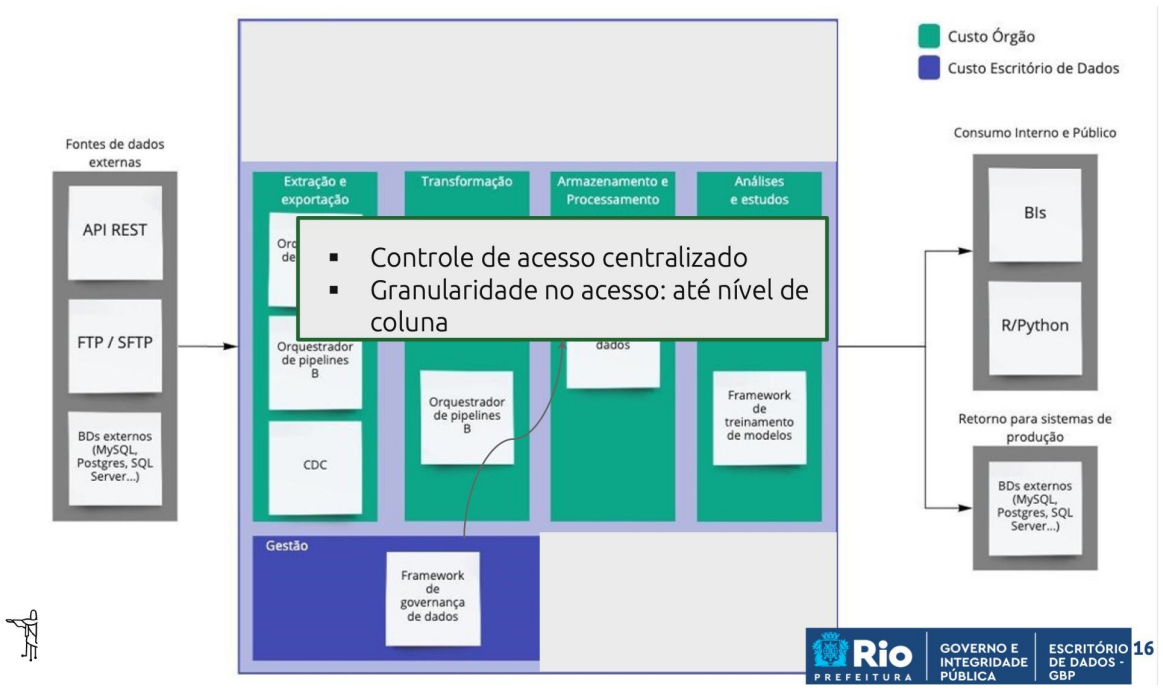

Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Transformação
Orquestrador de pipelines A
Orquestrador de pipelines B
CDC
Orquestrador de pipelines B
Consumo Interno e Público
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Transformações simples
No-code

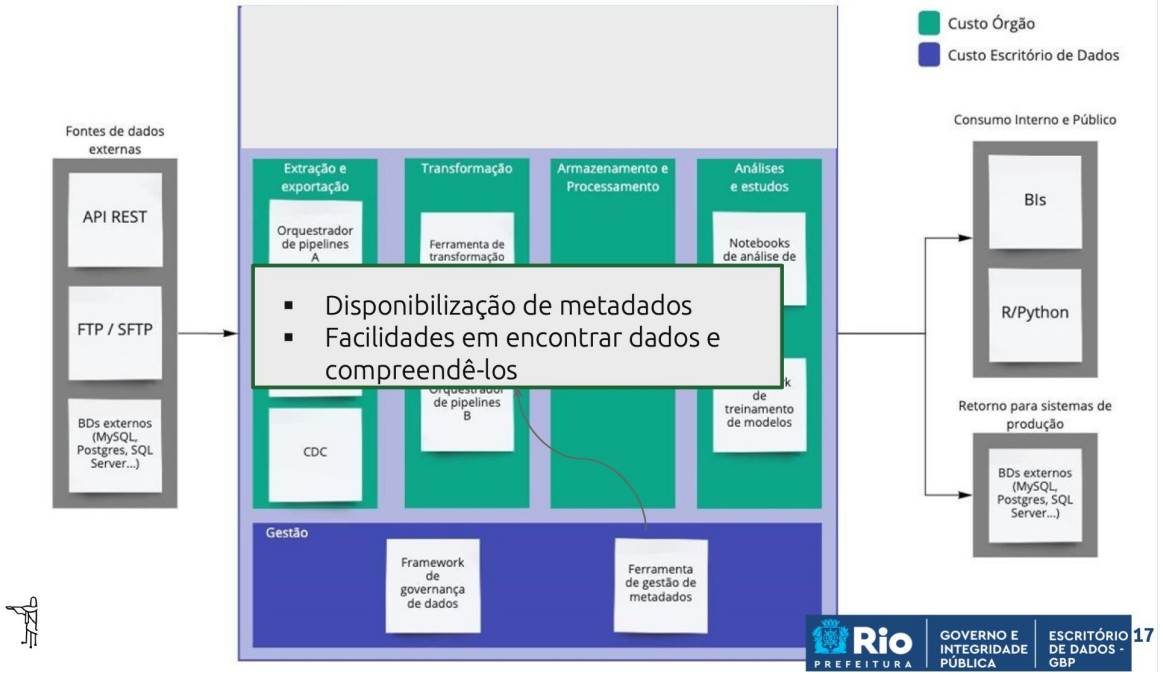

Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines A
Orquestrador de pipelines B
CDC
Transformação
Ferramenta de transformação de dados
Orquestrador de pipelines B
Consumo Interno e Público
BIs
R/Python
Qualquer tipo de transformação
Materialização de views
Linguagem SQL
Versionamento e execução descentralizados

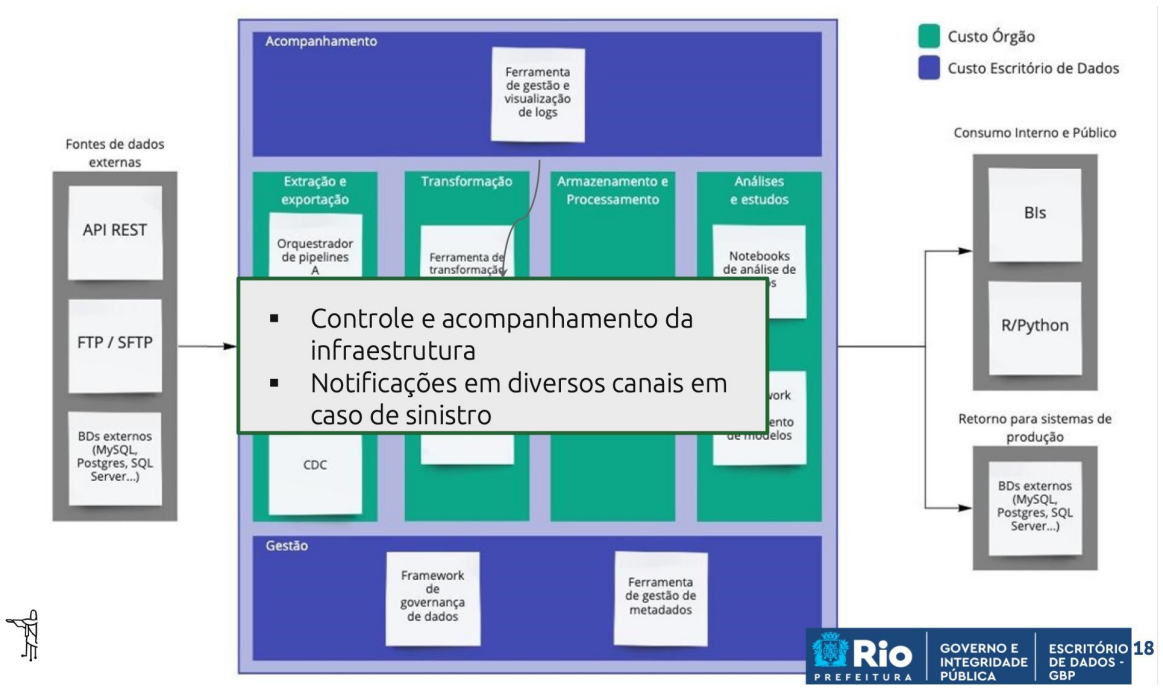

Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines A
Orquestrador de pipelines B
CDC
Transformação
Ferramenta de transformação de dados
Orquestrador de pipelines B
Armazenamento e Processamento
Armazém de dados
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Processamento de dados em escala de petabytes
API para interação via scripts
Custo descentralizado

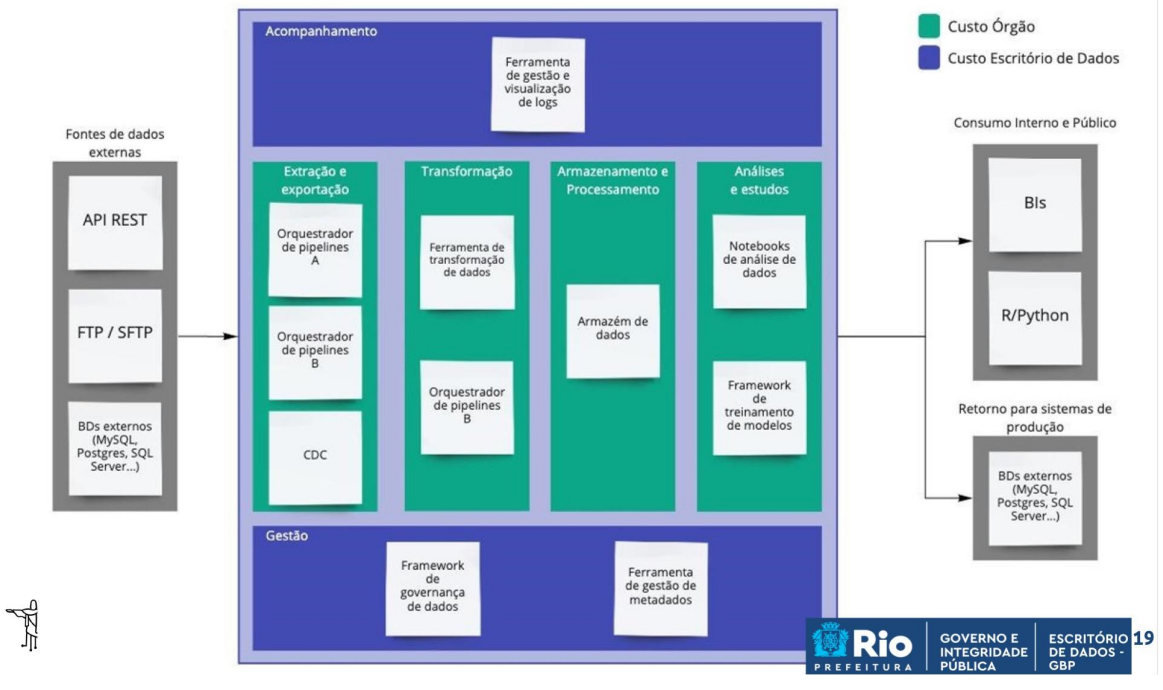

Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines A
Orquestrador de pipelines B
CDC
Transformação
Ferramenta de transformação de dados
Orquestrador de pipelines B
Armazenamento e Processamento
Armazém de dados
Análises e estudos
Notebooks de análise de dados
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Análises personalizadas
Real-time collaboration
Linguagens R, Python, Octave e Julia

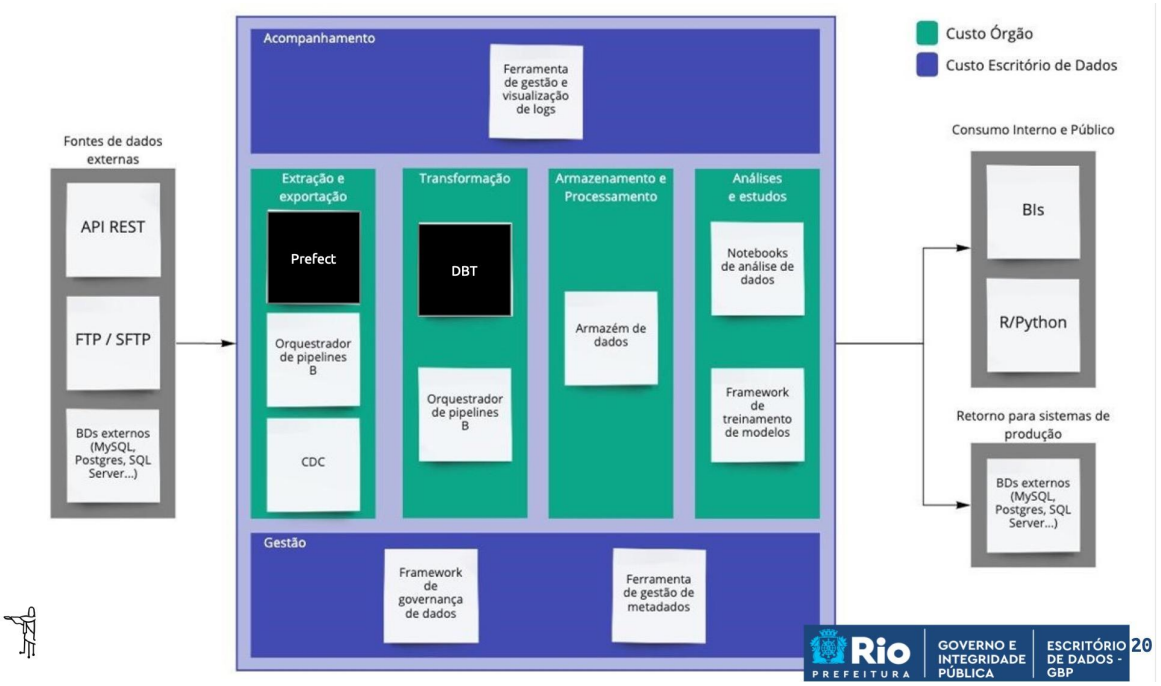

Custo Órgão
Custo Escritório de Dados
Treinamento de modelos em escala
MLOps
Serving
API RES
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Orquestrador de pipelines A
Orquestrador de pipelines B
CDC
Ferramenta de transformação de dados
Orquestrador de pipelines B
Armazém de dados
Análises e estudos
Notebooks de análise de dados
Framework de treinamento de modelos
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)


Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Transformação
Armazenamento e Processamento
Análises e estudos
Orquestrador de pipelines B
CDC
Orquestrador de pipelines B
Framework de treinamento de modelos
Gestão
Framework de governança de dados
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Controle de acesso centralizado
Granularidade no acesso: até nível de coluna
Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP
16


Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Transformação
Armazenamento e Processamento
Análises e estudos
Orquestrador de pipelines A
Ferramenta de transformação
Notebooks de análise de
CDC
Orquestrador de pipelines B
de treinamento de modelos
Gestão
Framework de governança de dados
Ferramenta de gestão de metadados
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Disponibilização de metadados
Facilidades em encontrar dados e compreendê-los
Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP
17

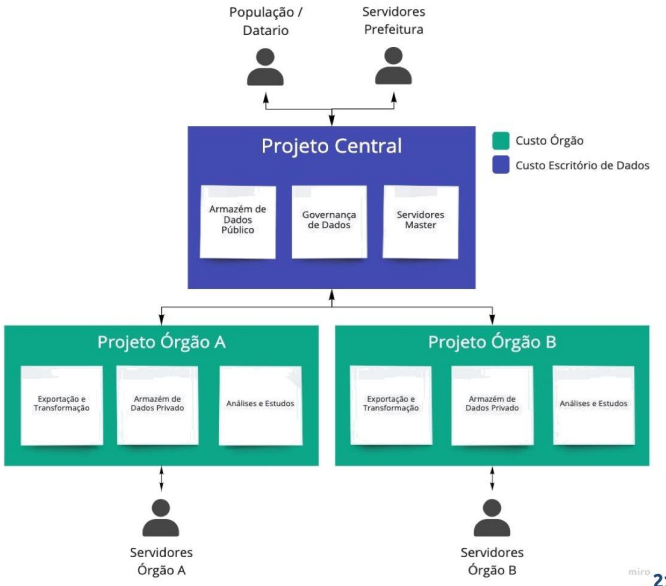

Acompanhamento
Ferramenta de gestão e visualização de logs
Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines A
CDC
Transformação
Ferramenta de transformação
Armazenamento e Processamento
Análises e estudos
Notebooks de análise de
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Gestão
Framework de governança de dados
Ferramenta de gestão de metadados
Controle e acompanhamento da infraestrutura
Notificações em diversos canais em caso de sinistro
Rio PREFEITURA
GOVERNO E INTEGRIDADE PÚBLICA
ESCRITÓRIO DE DADOS - GBP
18


Acompanhamento
Ferramenta de gestão e visualização de logs
Custo Órgão
Custo Escritório de Dados
Fontes de dados externas
API REST
FTP / SFTP
BDs externos (MySQL, Postgres, SQL Server...)
Extração e exportação
Orquestrador de pipelines A
Orquestrador de pipelines B
CDC
Transformação
Ferramenta de transformação de dados
Orquestrador de pipelines B
Armazenamento e Processamento
Armazém de dados
Análises e estudos
Notebooks de análise de dados
Framework de treinamento de modelos
Consumo Interno e Público
BIs
R/Python
Retorno para sistemas de produção
BDs externos (MySQL, Postgres, SQL Server...)
Gestão
Framework de governança de dados
Ferramenta de gestão de metadados
Rio PREFEITURA
GOVERNO E INTEGRIDADE PÚBLICA
ESCRITÓRIO DE DADOS - GBP
19



Coordenação de componentes

- Custos de armazenamento, processamento e consulta devem ficar em uma conta ligada ao órgão
- Custos de ferramentas de gestão devem ficar ligados ao Escritório de Dados e gestores centrais
- Cada órgão deve ser capaz de gerenciar acesso aos seus dados
- Gestores centrais devem ser capazes de gerenciar todos os acessos



Casos de uso

- Requisição de dados via API a cada minuto e processamento geográfico das informações maiores que 100GB
- Extração de dados de um banco de dados relacional com frequência diária e criação de algoritmos de ML com resultado sendo salvo em outro banco de dados relacional
- Análise de dados temporais e geoespaciais de telefonia móvel >1TB
- Obtenção e análise de dados em imagem e vídeo

Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP

Próximos Passos

- Publicação ainda em dezembro de 2021, da Consulta Pública com esta apresentação e catálogo de requisitos para que sejam avaliados pelos fabricantes de tecnologia alinhados com o tema.
- Envio de resposta dos principais fabricantes com sua visão da arquitetura de cloud com custos calculados até 10/01/2022
  - Se possível, apresentar arquitetura com todos os componentes da empresa e arquitetura mista com componentes Open-Source sugeridos

Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP

Anexo I – Requisitos (link aqui)

Função	Descrição	Relevância (maior, mais relevante)	Componentes Open-Source sugeridos	Componentes Open-Source (Exemplo)	Fatores para avaliação	A solução atende ao requisito?	Estimativa de carga	Proposta de componentes 1 (Exclusivamente provedor)	Estimativa de custo 1 (Exclusivamente provedor)	Comentários 1	Proposta de componentes 2 (Provedor + Compontes Open Source sugeridos)	Estimativa de custo 2 (Provedor + Compontes Open Source sugeridos)	Comentários 2
COPIE ESSA PLANILHA PARA PREENCHER													
Captura de dados alterados (CDC)	Rastrear alterações de dados na origem, replicando-as para o armazém de dados, mantendo os dados sincronizados	2	-	Debezium	Não deve criar triggers adicionais na origem Deve integrar com Armazém de Dados Diversas fontes (mínimo PostgreSQL, SQL Server, MySQL) Não deve submeter consultas ao SGBD		Backlog: 1TB Atualização: 100GB/mês						
Orquestrador de Pipelines A	Desenvolvimento de ETLs através da linguagem Python	3	Prefect	Prefect	Capacidade de uso pleno da linguagem Python: a ferramenta ou framework não pode impor qualquer limitação ao código Minimalismo: a transição de um Python script qualquer para uma pipeline em produção deve envolver poucas alterações de código Capacidade de agendamento de execuções Execução descentralizada (separação dos custos de computação) Capacidade de executar múltiplas pipelines a cada minuto Capacidade de obter log detalhado das pipelines Capacidade de integração com Slack, Discord, Telegram Interface web madura e intuitiva para acompanhar execução das pipelines Gestão de acesso da interface web para criação de grupos de usuários Autenticação de usuários para interface web Gestão de pipelines por projetos/temas		Início: 5 pipelines tipo A (a cada 1 min, 10 passos, ~0.5s/passo, ~0.5GB RAM) 10 pipelines tipo B (a cada 24 horas, 15 passos, 3s/passo, ~1GB RAM) 2 pipelines tipo C (a cada 24 horas, 5 passos, 15min/passo, ~8GB RAM) 3 pipelines tipo D (a cada 1 hora, 10 passos, 3s/passo, ~0.5GB RAM) Média: ~13h CPU/dia Perspectiva (6 meses): 20 pipelines tipo A 30 pipelines tipo B						
Orquestrador de Pipelines B	Desenvolvimento de ETLs através de uma interface gráfica	2	-	CDAP / Apache NiFi	Capacidade de desenvolvimento sem qualquer entendimento de programação Integração com diversas fontes de dados (REST APIs, bancos de dados, etc.), inclusive o próprio data lake Capacidade de agendamento de execuções Gestão de acesso aos usuários Gestão das pipelines por projeto/tema Execução descentralizada (separação dos custos de computação)		Início: 3 pipelines tipo B 2 pipelines tipo C Média: ~2,5h CPU/dia Perspectiva (6 meses): 10 pipelines tipo B 20 pipelines tipo C 10 pipelines tipo D Média: ~27h CPU/dia						
Ferramenta de transformação de dados	Versionamento e execução de pipelines de dados em linguagem SQL	3	DBT	DBT	Capacidade de verificação de dependência entre tabelas Diferentes tipos de materialização (mínimo: tabela nativa, view ou materialização incremental) Parametrização de queries Separação de schemas		Início: 5 execuções tipo A (a cada 1 hora, duração máxima ~3min, 5GB processados) 10 execuções tipo B (a cada 24 horas, duração máxima ~50min, 25GB processados) Perspectiva (6 meses): 30 execuções tipo A 30 execuções tipo B 10 execuções tipo C (a cada 24 horas, duração máxima ~2h, 50GB processados)						
Armazém de dados	Armazenamento e processamento de dados em larga escala	4	-	Ecossistema Hadoop	Capacidade de processamento e armazenamento ilimitada Otimização de processamento/custos via particionamento e correlatos Interação com o banco via SQL Funções de análise básica (SQL https://www.w3schools.com/sql/) Funções de análise de texto com regex Funções de análise geoespacial Conexões com plataformas de desenvolvimento de dashboards (PowerBI, Apache Superset) Funções de análise temporal (moving window, timezone) Suporte para json/array Gerenciamento automático e rápido de recursos para qualquer tipo de processamento Custos de processamento de responsabilidade do usuário Gestão de permissão de acesso aos dados para datasets/tabelas Gestão de permissão de acesso aos dados para datasets/tabelas/colunas ACID compliant Possibilidade de permitir que público acesse os dados Gestão de permissão de acesso por grupo de usuário Gestão de permissão de acesso por grupo de usuário		Início: processamento diario: 1TB armazenamento: 1TB Perspectiva (6 meses): processamento diario: 5TB armazenamento : 5TB						
Ferramenta de gestão e visualização de logs	Armazenamento e acompanhamento de logs de diversos serviços e aplicações	2	-	Elastic Stack	Capacidade de capturar logs de diversas fontes Diferentes níveis de mensagem (debug, informação, aviso, erro, etc.) Capacidade de redimensionamento automático (caso necessário) para atender a grandes rajadas de mensagens sazonais Facilidade de identificação de origem das mensagens		Início: 50 mensagens/minuto rajada de 50 mensagens a cada 1 hora rajada de 90 mensagens a cada 24 horas Perspectiva: 200 mensagens/minuto rajada de 250 mensagens a cada 1 hora rajada de 1500 mensagens a cada 24 horas						
Notebooks de análise de dados	Desenvolvimento de análises de dados interativas e replicáveis	1	-	CoCalc	Colaboração em tempo real Suporte a linguagens Python, Julia e R Capacidade de dimensionar as instâncias de execução de kernels Suporte a GPUs Suporte a formatação Markdown e LaTeX Suporte a adição de comentários no notebook (revisão) Versionamento e "blame"		Início: 16h/dia kernel tipo A (4 CPUs, 8GB RAM) 8h/dia kernel tipo B (8 CPUs, 32GB RAM) Perspectiva (6 meses): 96h/dia kernel tipo A 24h/dia kernel tipo B 4h/dia kernel tipo C (40 CPUs, 128GB RAM, 1 GPU)						
Framework de treinamento de	Pipelines de treinamento e serving	1	-	ML flow / Seldon Core	Implementação de pipelines em Python Suporte a diversos frameworks (mínimo Tensorflow, Torch e Scikit Learn) Disponibilização de modelos para inferência via HTTP REST API		Início: - Treinamento: 24h/mês instância tipo A (8 CPUs, 32GB RAM, 1 GPU) - Serving (inferência via HTTP): N/A Perspectiva (6 meses):						

modelos	de modelos em escala				Capacidade de dimensionar as instâncias de treinamento e serving (inferência) Suporte a redimensionamento automático de instâncias de serving Suporte a GPUs	- Treinamento: 96h/mês instância tipo A 72h/mês instância tipo B (16 CPUs, 64GB RAM, 2 GPUs) - Serving (inferência via HTTP): 24h/dia de 3 modelos da ordem de grandeza da EfficientNetB6 (~45M parâmetros)	
Framework de governança de dados	Ferramenta capaz de gerir acesso aos dados e acompanhar linhagem dos dados	3	-	Apache Atlas	Integração com o armazém de dados Interface web Gestão de permissão dos acessos por usuário Gestão de permissão dos acessos por grupo de usuário Granularidade de acesso a nível de coluna em tabelas do armazém de dados Capacidade de disponibilização pública dos dados Versionamento e "blame"	Início: 8 administradores 32 usuários Acesso público 60 pessoas/mês Perspectiva (6 meses): 24 administradores 72 usuários Acesso público 500 pessoas/mês	
Ferramenta de gestão de metadados	Ferramenta para gestão e busca de metadados	1	-	Apache Atlas	Integração com o armazém de dados Interface web Geração automática de documentação Funcionalidades de busca de dados através de metadados Versionamento e "blame"	Início: Perspectiva (6 meses):	
Geral	Requisitos gerais da arquitetura	4	-	-	Número total de componentes da solução Solução em Cloud Disponibilidade de servidores no Brasil Número de datacenters no Brasil Certificado ISO 27701 e 27001 para todos os componentes		
Observações Gerais + Custo não relacionado a componentes							
						Custo Total Estimado:	

Rio PREFEITURA | GOVERNO E INTEGRIDADE PÚBLICA | ESCRITÓRIO DE DADOS - GBP

Consulta Pública para definição de Arquitetura em Nuvem para Projetos de Dados

Rio de Janeiro - Dez/2021

Anexo 2 - Tabela comparativa das respostas