serpro.gov.br

Consulta Pública Eletrônica para Contratação de Solução de Integração de Dados com Autosserviço e Monitoramento.

DIOPE/SUPEC/ECTAN
05/2020

SERPRO

Sumário

1. Objeto

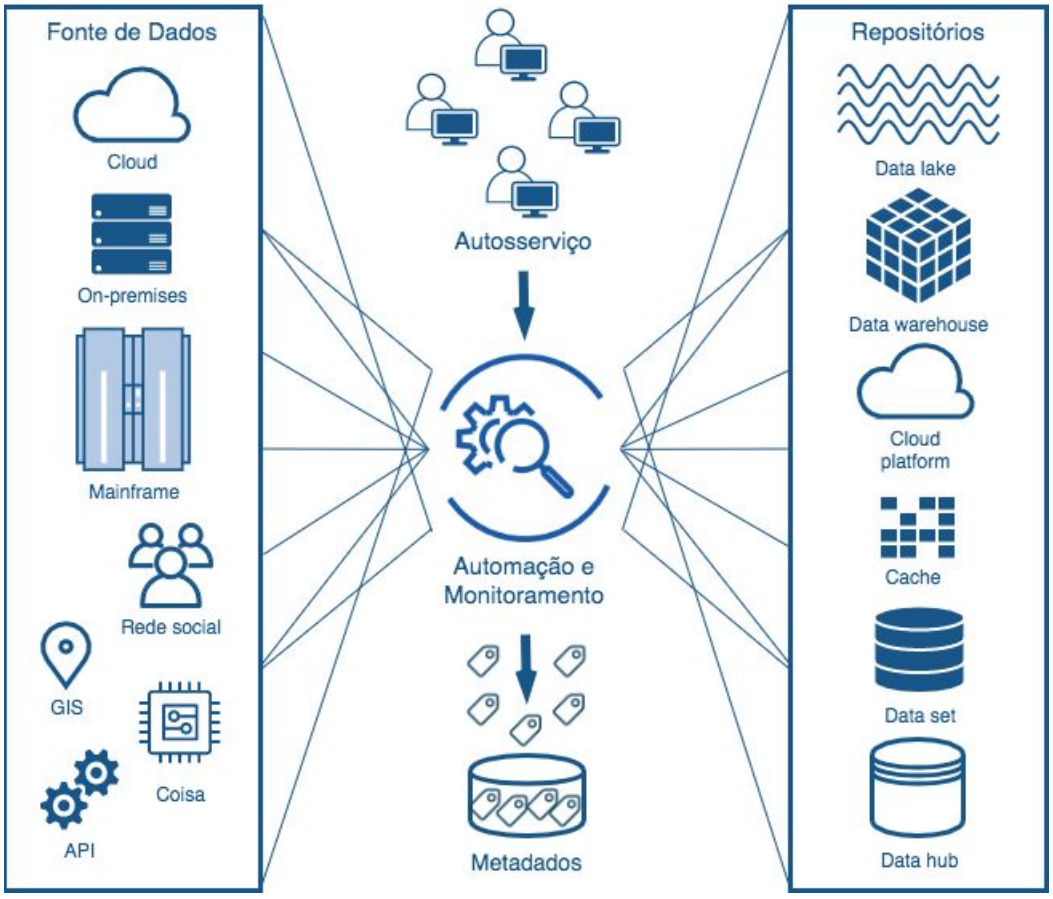
1.1. Consulta pública eletrônica para subsidiar decisão futura sobre contratação de solução de integração de dados, com autosserviço e monitoramento, a ser inserida na arquitetura de dados do Serpro.

2. Especificação do Objeto

2.1. Solução de integração de dados, com autosserviço e monitoramento, a ser inserida na arquitetura de dados do Serpro, de modo a sustentar a orquestração de fluxos de dados para as diversas estruturas data lake, data warehouse, cloud platform, cache e data set mantidas pelo Serpro.

2.2. O escopo da consulta pública eletrônica refere-se apenas as funcionalidades de aquisição de dados, replicação de dados, integração de dados, preparação de dados, virtualização de dados, mascaramento de dados e criação e gerenciamento de fluxos de dados.

2.2.1. Arquitetura Conceitual de Dados



2.2.1.1. Com esta arquitetura, o Serpro deseja modernizar seu gerenciamento de dados e padrões arquitetônicos que, quando combinados, possam suportar cargas de trabalho cada vez mais complexas, diversas e distribuídas e possam permitir a integração, o compartilhamento e o controle dos dados, além de garantir alinhamento da abordagem com os requisitos de negócios, combinando as características e os casos de uso comuns de cada uma dessas estruturas data lake, data warehouse, cloud platform, cache e data set.

2.3. Às empresas interessadas devem responder a consulta pública com as seguintes informações:

2.3.1. **Contatos:**

2.3.1.1. Nome completo do responsável pelas respostas desta Consulta Pública.

2.3.1.2. Cargo, telefones e endereço de e-mail.

2.3.2. **Identificação da Empresa:**

2.3.2.1. Nome completo e fantasia.

2.3.2.2. CNPJ.

2.3.2.3. Endereço completo.

2.3.2.4. Site WEB (www).

2.3.3. **Solução**

2.3.3.1. Nome da solução oferecida, objeto desta consulta pública.

2.3.2.2. Site WEB do fabricante da solução (www).

2.3.3.3. Descrição detalhada da solução e seus componentes (Documentos/datasheet, etc).

2.3.3.4. Forma de licenciamento da solução e seus componentes (Licença perpétua /subscrição anual, e outras), conforme exemplo abaixo:

Part Number	**Descrição da Solução**	**Licença de uso perpétuo / subscrição anual**	**Unidade / Métrica**	**Faixa / Quantidade**	**Estudo de referência do Valor Unitário (R$)**

2.3.3.5. Forma e condições de pagamento da solução e seus componentes (Licença perpétua /subscrição anual, etc).

2.3.4. **Base de Clientes:**

2.3.4.1. Quantidade de clientes no Brasil.

2.3.4.2. Nomes dos entes públicos que já adquiriram a solução.

2.3.5. **Experiência e Suporte:**

2.3.5.1. Possui equipe de suporte técnico para atendimento fora do horário comercial e em dias não úteis.

2.3.5.2. O suporte é prestado pelo fabricante ou parceiro?

2.3.5.3. Quais os níveis de serviços ofertados para a solução (Tempo de atendimento, tempo de solução, etc).

2.3.5.4. Informar a forma de repasse de conhecimento, resumos das grades e carga horária.

2.4. Os requisitos funcionais e não funcionais da solução estão descritos no documento "Anexo A - Requisitos Solução Integração de Dados".

2.4.1. Orientações para preenchimento do anexo A.

2.4.1.1. Os campos a serem preenchidos nesta planilha estão marcados em azul:

2.4.1.2. O escopo para o preenchimento das respostas é dividido em campos objetivos e descritivos.

2.4.1.3. Para os campos objetivos, utilize o seguinte padrão de respostas:

2.4.1.3.1. "0" = requisito NÃO atendido pela solução.

2.4.1.3.2. "1" = SIM, requisito totalmente atendido pela solução.

2.4.1.3.1. "0,5" = requisito PARCIALMENTE atendido pela solução.

2.4.1.4. Se não atende, como poderia atender? Indique customizações necessárias e de baixa complexidade para ser atendido plenamente - a empresa se compromete a implementá-lo na totalidade.

2.4.2. Requisitos Funcionais:

Item	Requisitos Funcionais	Detalhamento dos Requisitos	Forma de Atendimento	
			Atende? 0: Não 0,5: Parcial 1: Sim	Se não atende, como poderia atender?
1	Fonte de Dados	Conectividade à fonte de dados relacionais, no mínimo Oracle, Sql Server, PostGreSQL, DB2, My SQL e Maria DB.		
2		Conectividade à fonte de dados mainframe, no mínimo Adabas, DB2 e Vsam.		

3		Conectividade a fonte de dados não relacionais (NoSQL), no mínimo Elasticsearch, Hadoop, Hbase, SolR, MongoDB, Cassandra, Redis e Neo4J.		
4		Conectividade a fonte de dados em nuvem SaaS ERP, CRM (complementar SalesForce, SAP...)		
5		Conectividade à fonte de dados em vários formatos de arquivo (como XML, JSON, CSV, TXT, PDF Avro, ORC, RCFile e Parquet)		
6		Conectividade à fontes de dados disponíveis nas tecnologias e/ou protocolos JDBC, ODBC, HDFS (block size padrão 64MB), S3, NFS, SAN, LDAP, HTTP/HTTPS, FTP e SFTP.		
7		Conectividade a redes sociais, estando incluso no mínimo Facebook, Twitter, LinkedIN e Instagram;		
8		Conectividade à infraestrutura de mensageria, incluindo aquelas providas por middleware de integração, no mínimo Kafka, RabbitMQ, IBM MQ, TIBCO EMS, JMS e MQTT.		
9		Conectividade a ambientes baseados em Apache Hadoop, seja como origem ou destino dos dados, sendo compatível, no mínimo, com as seguintes tecnologias HDFS, Apache Hive, Apache Impala, HBase, Spark, Apache Kafka, Apache Kudu e Apache Solr.		
10		Conectividade a séries temporais e dados geoespaciais residentes em sistemas como GIS, ESRI e outros.		
11		Conectividade com soluções proprietárias, de provedores de nuvem, Azure SQL, Azure Cosmos DB, Amazon RDS, Amazon Redshift, Amazon S3, DynamoDB NoSQL, ElastiCache, Neptune graph, Snowflake, Glacier, services Google.		
12		Conectividade à dados de streaming via		

		plataformas ESP (Event Stream Processing) tal como IoT, log, dados de sensores, etc.		
13	Aquisição de Dados	Execução de tarefas bulk e/ou batch de extração de dados e abordagem de entregas ETL/ELT/ETLT.		
14		Execução de tarefas de carga de dados usando SQL ou qualquer outra linguagem nativa do SGBD.		
15		Execução de tarefas de carga de dados a partir de consultas a componentes Hadoop, usando o modelo MapReduce, Spark ou linguagens ou mesmo construções sobre este modelo.		
16		Execução de tarefas de transferência de arquivos (MFT) entre sistemas, com recursos para alta taxa de transferência, compactação automática, reinicialização do ponto de verificação e segurança, como criptografia.		
17		Execução de tarefas CDC (Change Data Capture), incremental, identificando automaticamente novos registros, alterações ou exclusões de registros já existentes. Esse tipo de aquisição deve ocorrer sem prejuízo ao desempenho do ambiente produtivo, ou seja, não deve: submeter consultas ao SGBD, necessitar de qualquer coluna identificadora ou criar triggers adicionais.		
18		Execução de tarefas CDC para as fontes de dados relacionadas nos requisitos de 1 a 11.		
19		Execução de streaming de dados a partir de fontes internas, tais como websites, sensores, dispositivos e aplicações e fontes externas, tais como redes sociais e plataformas ESP (ver requisitos 7 e 12).		

20		Suporte a dois tipos de cargas: - Completa, através da execução de consultas no banco de dados de origem, usando SQL ou qualquer outra linguagem nativa do SGBD e execução de consultas aos componentes Hadoop, usando os modelos MapReduce e Spark ou linguagens ou mesmo construções sobre estes modelos; - Incremental, identificando automaticamente novos registros, alterações ou exclusões de registros já existentes. Esse tipo de carga deve ocorrer sem prejuízo ao desempenho do ambiente produtivo, ou seja, não deve: submeter consultas ao SGBD, necessitar de qualquer coluna identificadora ou criar triggers adicionais no banco de dados e não deve submeter consultas aos componentes do Hadoop. Recomenda-se o uso de recursos como o CDC (Changed Data Capture) baseado em log de transação.		
21		Ter métricas para verificação de consumo de CPU, Memória, volume de tráfego de rede, frequência de execução de tarefa (job/fluxo/pipeline). Tais métricas devem ser aferidas na origem, na própria ferramenta e no destino.		
22	Replicação de Dados	Execução de tarefas para cópia dos dados, movendo-as fisicamente, quase em tempo real, de um local para outro, sempre em um repositório de dados físico, sem alterar a estrutura ou o conteúdo dos dados que são movidos		
23		Execução de tarefas de sincronização de DML (inclusões, alterações e exclusões de registros de dados)		
24		Execução de tarefas de sincronização de DDL (inclusões, alterações e exclusões nos objetos do banco de dados, sejam eles colunas em tabelas, stored procedures ou triggers)		

25		Sincronização unidirecional e/ou bidirecional, a critério do usuário.		
26		Ter métricas para verificação de consumo de CPU, Memória, volume de tráfego de rede, frequência de execução de tarefa (job/fluxo/pipeline). Tais métricas devem ser aferidas na origem, na própria ferramenta e no destino.		
27	Replicação de Dados entre Plataformas Alta - Baixa	Execução de tarefas de cópia de dados entre as plataformas alta e baixa, com CDC (Changed Data Capture) baseado em log de transação (ver requisito 2).		
28		Possibilidade de otimização de peformance e redução de consumo de recursos na plataforma mainframe. As métricas para verificação são MIPS e MSU.		
29	Integração de Dados	Criação de fluxo de dados para mashup e mesclagem de dados; limpeza de dados; filtragem; e cálculos, grupos e hierarquias definidos pelo usuário, a partir de várias fontes de dados para criar novos datasets, ou atualizar datasets existentes.		
30		Identificação de relacionamentos entre vários objetos de dados. Isso inclui modelagem / estruturação de dados que permita aos usuários especificar tipos e novos relacionamentos de dados.		
31	Preparação de Dados	Crição de fluxo de dados que compreenda tarefas de padronização, validação, limpeza, enriquecimento, deduplicação e consolidação de dados, garantindo a entrega de informações de alta qualidade.		
32		Suporte para recursos que melhorem a transformação de dados por meio de aprendizado de máquina (ML).		
33		Compatíbilidade com campos do tipo texto, independente do tamanho; numéricos, sejam eles inteiros ou decimais, independente da precisão; data e hora, independente do formato e precisão, geográficos, binários.		

34		Suporte a tipos complexos (struct, array, map), em especial ao ingerir XML, JSON e banco de dados Adabas.		
35		Suporte a dados semiestruturados e não estruturados, como email, sites, ferramentas de escritório (Word, Excel), repositórios de conteúdo, arquivos de áudio e vídeo.		
36		Manipulação de tipos numéricos, como cálculos simples em campos numéricos, como soma, multiplicação, divisão, arredondamentos, etc.		
37		Manipulação de tipos texto, como substituição de caracteres, extração de trechos do texto, etc.		
38	Migração de fluxo de dados	Migração de forma transparente um fluxo preexistente entre os tipos de carga, sem retrabalho equivalente ao de criar novamente o fluxo.		
39	Schedule de fluxo de dados	Módulo para agendamentos de execução dos fluxos de carga (stream-based), independentemente do tipo (completa ou incremental), com periodicidade e em horários determinados pelo usuário, suprindo no mínimo, janelas do tipo: D-1 (1 dia de defasagem), H-2 (2 horas de defasagem) .		
40	Desenvolvimento de fluxo de dados	Módulo para desenvolvimento para os fluxos de carga (stream-based), contendo ambientes distintos (desenvolvimento e produção, por exemplo), incluindo aprovação e promoção entre ambientes.		
41		Funcionalidades para construção visual dos fluxos de dados através de recurso flow-based programming.		
42	Teste e depuração de fluxo de dados	Módulo para depuração de erros dos fluxos de carga (stream-based) desenvolvidos pela equipe.		

43	Resiliência em fluxo de dados	Continuidade do fluxo de dados, caso ocorra indisponibilidade temporária durante o processo, sem a necessidade de intervenção do usuário, e sem a ocorrência de perda ou duplicação de informações, inclusive quando este processamento for distribuído.		
44	Gerenciamento de fluxo de dados	Funcionalidades para acompanhamento de diferentes tipos e tamanhos de fluxos de carga, controle de versão dos fluxos de carga que permita colaboração e auditoria, contemplando, no mínimo, histórico de agendamento dos jobs de carga, histórico de execução desses jobs de com estatísticas da execução (no mínimo, tempo de execução e número de registros afetados).		
45		Monitoramento de fluxo de dados em execução, indicando servidor, banco de dados e consulta, permitindo parar, cancelar e retomar o processamento do fluxo, e alertas em casos de falha nos fluxos de carga (e-mail, mensagens etc.)		
46		Recursos de rastreamento constante do comportamento e do uso dos recursos dos fluxos de carga.		
47	Bulk data	Funcionalidade para importar e exportar em massa (bulk data) na origem ou destino dos dados, a ferramenta deve fazer uso desse recurso.		
48	Schema Registry	Uso do Schema Registry (estrutura para criar repositórios de metadados) para fornecer uma camada de veiculação de metadados, com interface RESTful para armazenar e recuperar esquemas.		
49	Repositório de controle de versão	Uso do Git como repositório de controle de versão para o código de integração de dados		
50	Integração contínua	Continuous Integration/Continuous Development (CI/CD)		

51	Suporte ativo a metadados	Descoberta de metadados aprimorada pelo aprendizado de máquina e análises internas para suportar, otimizar e até automatizar tarefas de gerenciamento e integração de dados humanos.		
52	Suporte passivo a metadados	Suporte metadados passivos, incluindo aquisição automatizada de metadados, dados, criação de modelo, documentação e manutenção, análise de linhagem e impacto, um repositório de metadados aberto e sincronização de metadados com uma interface do usuário final para visualizar e trabalhar com metadados.		
53	Sincronização de Metadados	Sincronização unidirecional e/ou bidirecional metadados, no mínimo, descrição de tabela e descrição de colunas		
54	Importação e Exportação de Metadados em Massa	Funcionalidade para importar e exportar metadados, com interface RESTful para armazenar e recuperar metadados.		
55	Virtualização de dados	Virtualização dados, execute consultas em fontes de dados distribuídas para criar visualizações virtuais integradas de dados "na memória" e forneça resultados de várias maneiras para o consumo downstream.		
56		Virtualização de dados incluindo dados residentes na plataforma alta (ver requisito 2).		
57		No caso de virtualização de dados incluindo dados residentes na plataforma alta, ter a possibilidade de otimização de peformance e redução de consumo de recursos na plataforma mainframe. As métricas para verificação na plataforma alta são MIPS e MSU. As métricas na plataforma baixa são consumo de CPU, Memória, volume de tráfego de rede, frequência de execução de tarefa (job/fluxo/pipeline). Tais métricas devem ser aferidas na origem, na própria ferramenta e no destino.		

58	Escalabilidade / desempenho	Forneça taxa de transferência e tempos de resposta adequados para satisfazer os SLAs de desempenho para todos os casos de uso de integração de dados e todos os requisitos de granularidade e latência de dados, dados os volumes cada vez maiores de dados e as necessidades de diversidade de dados.		
59	Suporte a mascaramento persistente de dados	Funcionalidade para Identificar dados confidenciais em colunas, em arquivos de formato não estruturados (PDF, arquivos do Excel, arquivos de texto, arquivos de log etc.) e conteúdos semiestruturados (XML, JSON etc.) .		
60		Funcionalidade para proteger os dados confidenciais sem prejudicar o projeto, por meio da remoção, mascaramento ou simplificação dos dados considerados confidenciais, especialmente quando os dados serão carregados em ambientes de desenvolvimento, testes, validação ou homologação.		
61		Funcionalidade para mascarar dados confidenciais com diversas formas e tipos, como nomes, data/hora, números, documentos pessoais, endereços de e-mail, CEP, cartão de crédito e tipo sanguíneo.		
62		Funcionalidade para incluir ou customizar conjunto de regras que definem os campos que serão mascarados e a função de mascaramento que será usada.		
63		Funcionalidade para permitir criação de formatos de mascaramento definidos pelo usuário		
64		Funcionalidade para criar banco de dados de teste (clone da origem) e, em seguida, importe os dados mascarados para o banco de dados de teste.		
65		Funcionalidade para mascarar de dados de atualização, isto é, ao ocorrer substituição de valor do atributo, tanto o novo valor do		

		atributo como o histórico, se mantido histórico, devem ser mascarados.		
66		Funcionalidade para usar armazenamento intermediário (Staging Area) no processo de mascaramento de dados persistente para auxiliar a transição dos dados da origem para o destino. Nessa área são tratadas as regras e padrões predeterminados de mascaramento para então prosseguir para a carga.		
67		Os dados mascarados devem ser realistas o suficiente para serem úteis para fins de desenvolvimento, testes e análise.		
68		Os dados mascarados devem ser irreversíveis, isto é, que não seja possível recuperar os dados originais a partir dos dados mascarados.		
69	Suporte a mascaramento dinâmico de dados	Em consultas realizadas via interfaces da solução, limitar a exposição de dados confidenciais através do mascaramento dinâmico dos dados para usuários sem privilégios.		
70		Permita integração via API a soluções de Data Loss Prevention		
71		Permita integração via API a soluções de Mascaramento Dinâmico de Dados		
72		O mascaramento em tempo real é aplicado quando um usuário acessa dados com base em privilégios de acesso.		

2.4.3. Requisitos Não Funcionais

Item	**Requisitos Funcionais**	**Detalhamento dos Requisitos**	**Forma de Atendimento**	
			Atende? 0: Não 0,5: Parcial 1: Sim	**Se não atende, como poderia atender?**

1	Interface com o usuário	Interfaces de usuário compatíveis com distribuições Ubuntu, Windows e MacOS em estações de trabalhos, navegadores Google Chrome, Firefox, Apple Safari e Microsoft Edge. A interface com o usuário em navegadores não poderá fazer uso de Applets Flash e Java.		
2		Interface gráfica (GUI – Graphical User Interface) e interface de linha de comando (CLI – Command Line Interface). A interface gráfica deve conter, no mínimo, Organização hierárquica por projetos, Possibilidade de reutilização de scripts, Objetos da origem e destino (bancos de dados, tabelas e colunas), Modelo de dados da origem e destino, Módulo único para criação de fluxos de carga (completo e incremental), Módulo único para monitoramento dos fluxos de carga (completo e incremental), Módulo único de histórico de execução dos fluxos de carga (completo e incremental), Interface gráfica adequada para uso por perfis de usuário com menor qualificação técnica, como analistas de negócio, permitindo que os mesmos criem seus fluxos de carga ou de qualidade de dados com independência da equipe de TI, incluindo interface a ser apresentada no idioma de escolha do usuário com documentação on-line.		
3		Possibilidade de customização de painéis para monitoramento de execução de tarefas e problemas.		
4	Integrações	API para utilização programática, permitindo que um desenvolvedor crie chamadas a esta API que possa, no mínimo: verificar o histórico de processamento dos fluxos de carga/qualidade, iniciar o processamento de um fluxo de carga/qualidade, monitorar os processamento em execução, parar a execução dos fluxos de carga/qualidade,		

		cancelar a execução dos fluxos de carga/qualidade, e retomar a execução dos fluxos de carga/qualidade.		
5		Integração com barramento de dados (ElasticSearch, Tibco), integração com ferramentas de monitoramento (Prometheus, Suite HPE, Suite Broadcom), integração com Analytics/BI, como PowerBI, Qlik e Tableau (entre vários outros), para que possam ser invocados diretamente nessas ferramentas pelos usuários.		
6		Fornecimento de bibliotecas de desenvolvimento de software (SDK) e permita que os usuários criem conectores adicionais para fontes de dados exclusivas.		
7	Segurança	Integração com Red Hat Directory Server 10, para autenticação de passagem (pass-through authentication) e autenticação baseada em perfis (role-based authentication).		
8		Log e auditoria, permitindo identificar, responsável por inclusão, alteração ou exclusão de fluxo de dados/qualidade		
9		Segurança granular baseada em perfis, permitindo associar funcionalidades da ferramenta a determinados perfis.		

3. Publicação

3.1. Consulta pública com fulcro no Art. 31, da Lei n° 9.784/1999, objetivando esclarecimentos sobre produtos, processos, soluções e tecnologias junto ao mercado.

4. Período

4.1. A consulta pública eletrônica ficará publicada pelo período de 8 (dias) úteis, podendo ser prorrogado a critério do SERPRO.

5. Responsáveis

4.1. A Consulta Pública Eletrônica será acompanhada pelos empregados:

4.1.1. Charles Morais Magalhães, matrícula 21065411, lotado na DIOPE/SUPEC/ECTAN, Telefone: (61) 2021-7259, e-mail: **charles.magalhaes@serpro.gov.br**.

4.1.2. Flavia Costa Bomfim, matrícula 01093649, SUPAI/AIGSB/AIGUP, e-mail: **flavia.bomfim@serpro.gov.br.**