Big Data, Data Science e Machine Learning com Google Cloud Platform

Data Science Academy

Capítulo 7
Cloud DataPrep

Introdução

Diferença entre ferramentas de preparação de dados tradicionais

Fluxos, receitas, steps, outputs, execução de jobs

Tratamentos de valores mismatched e missing

Transformações de dados

Transformações preditivas

Introdução

Transformações em arquivos JSON e arquivos de pares chave-valor

Lookups e Joins

Importação de Dados do BigQuery

Execução de jobs e exportação de resultados

Posso preparar meus dados utilizando R / Python / Tableau / PowerBI, etc?

Cloud DataPrep

Serviço para exploração visual, limpeza e preparação de dados estruturados e não estruturados para análise

O serviço é serverless e escalável, sem infraestrutura para gerenciar

Possibilita preparação de dados através de uma interface visual, sem escrita de código

O serviço é operado pela parceira Trifacta, que opera em conjunto com o Google para prover uma ótima experiência ao usuário

Cloud DataPrep

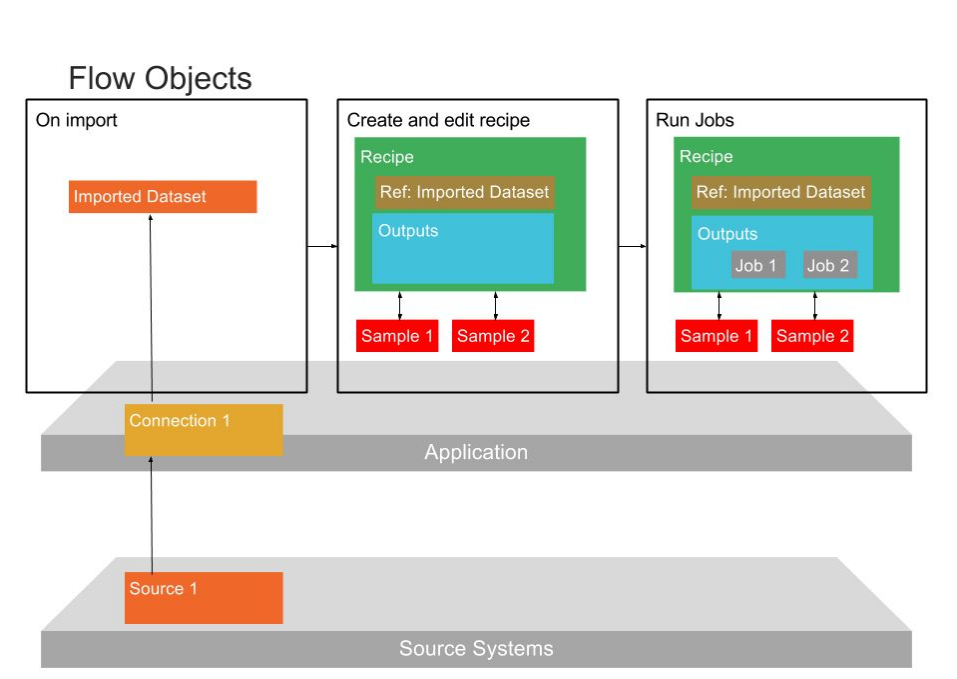
Automaticamente detecta schemas, tipos de dados, possíveis joins e anomalias como valores missing, outliers, valores duplicados, entre outros

Suporta variadas fontes de dados como CSV, JSON, tabelas relacionais e é capaz de escalar para terabytes

Construído no topo do Google Cloud DataFlow

Integrado com Google Cloud Platform, possibilitando acesso a serviços como Cloud Storage, BigQuery e gerenciamento de permissões através do IAM

DataPrep WorkFlow: Fluxos



Um fluxo possui um ou mais datasets importados, receitas e outros objetos

Um dataset importado é uma referência a “dados originais” - os dados em si não existem dentro do Dataprep

“Dados originais” podem ser uma referência a um arquivo, múltiplos arquivos, uma tabela de um banco de dados, etc

DataPrep WorkFlow: Receitas

Uma “receita” é uma sequência de passos que são aplicadas para transformar um dataset

Receitas serão interpretadas pelo DataPrep e se tornarão comandos que serão executados para transformação dos dados

Uma receita é criada a partir de um “dataset importado” ou a partir de outra receita

Importante: as etapas de uma receita não são imediatamente executadas sobre o dataset importado. Na realidade, as operações só serão efetivadas quando um “job” for executado

DataPrep WorkFlow: Saídas e Destinos de Publicação

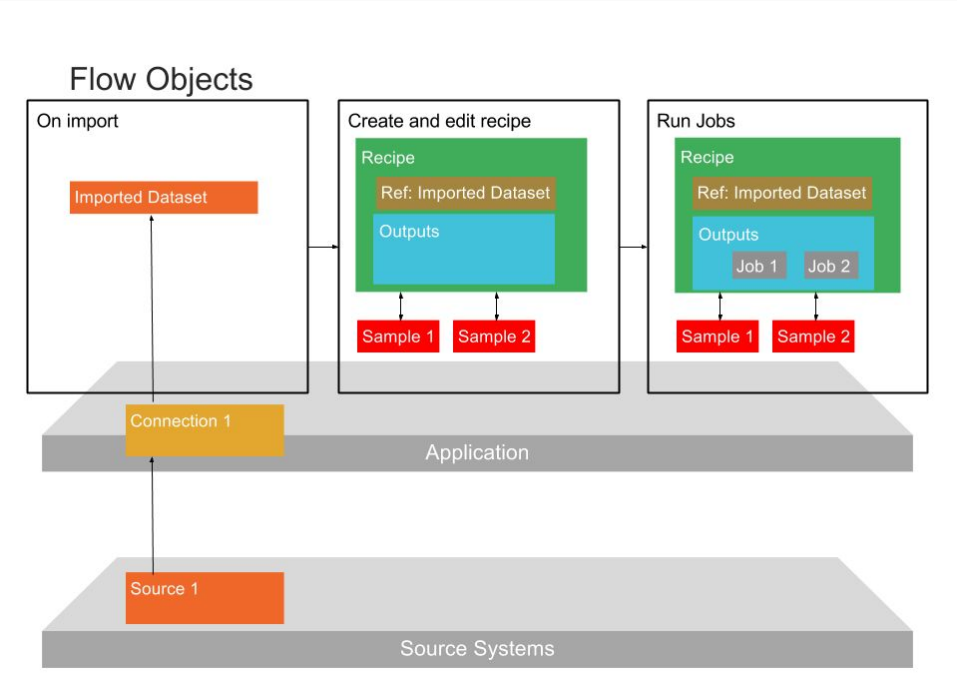
Um “output” contém 1 ou mais “publishing destinations”, que define o formato de saída, localização e outras opções que serão aplicadas aos resultados gerados a partir da execução de um job

Uma receita pode ter múltiplos destinos de publicação

DataPrep WorkFlow: Fluxos e receitas

Uma “window_frame_clause” define uma “window frame” na linha atual dentro de ur valiada.

Os



Obrigado

Data Science Academy