ESTUDO DE CASO MAGALU

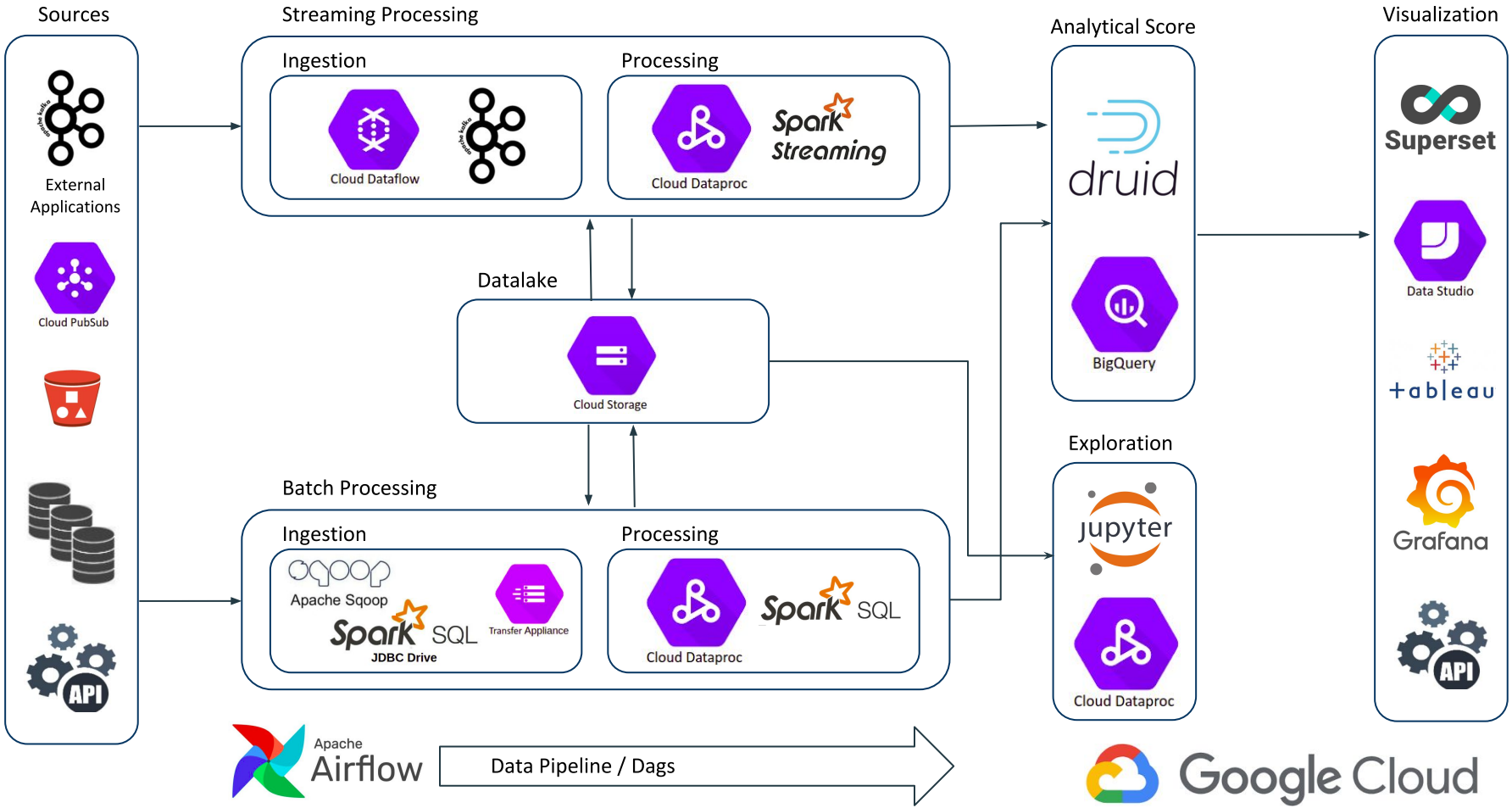
Plataforma de Dados

Uni-FACEF
Centro Universitário Municipal de Franca

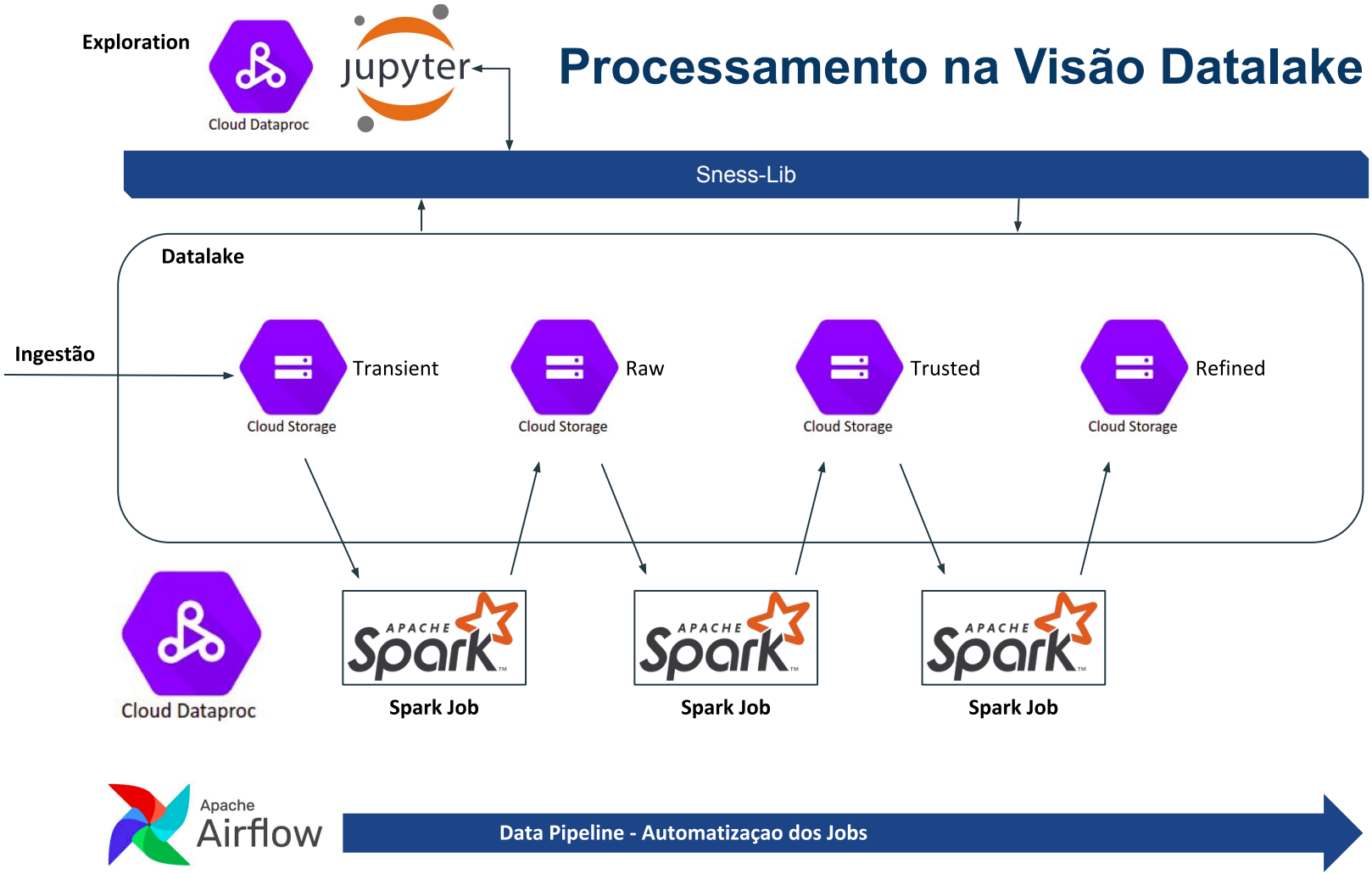
PÓS GRADUAÇÃO

Especialização Desenvolvimento de Aplicações Web e Móveis Escaláveis
Módulo: BIG DATA COM PYTHON - Professor: André Morais
Turma 2019-2020

Arquitetura Macro

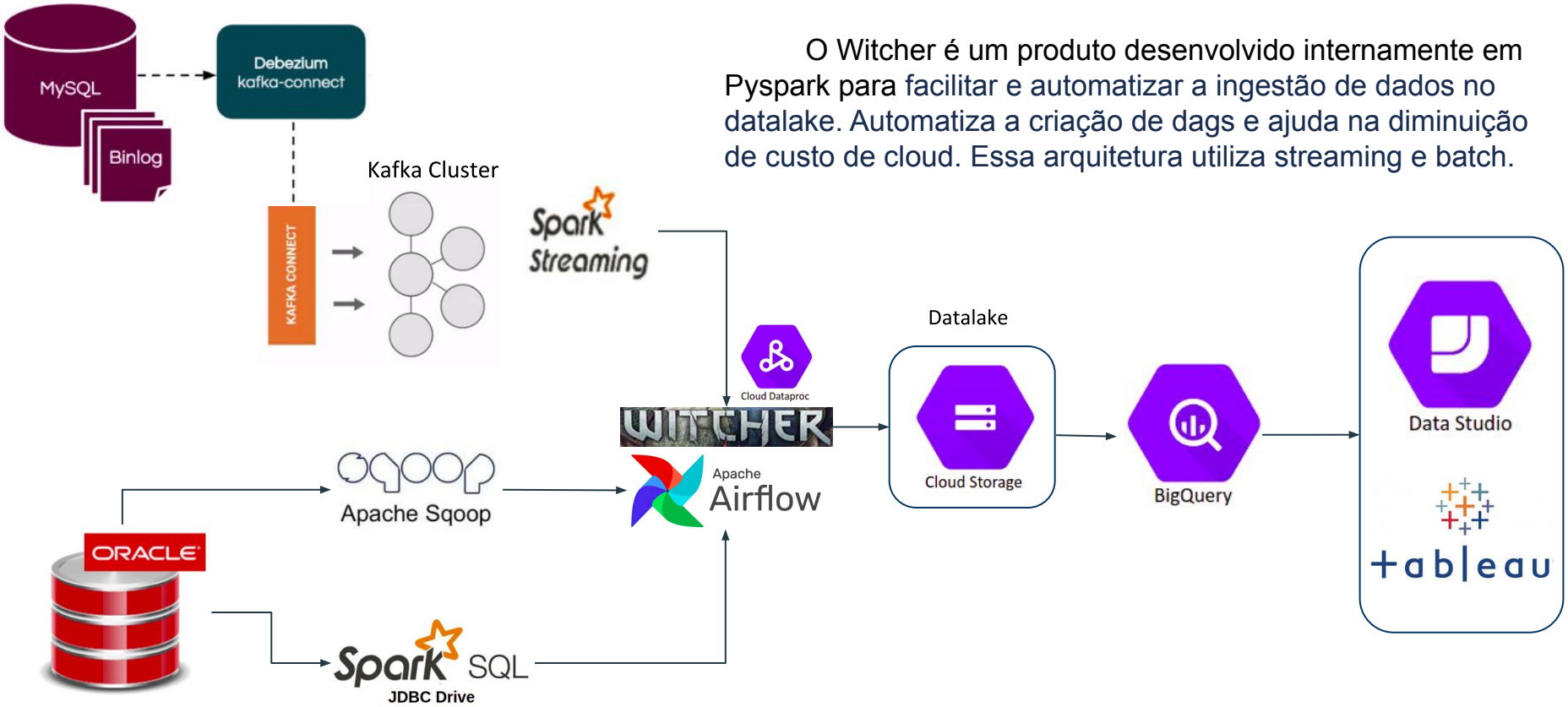


Processamento na Visão Datalake



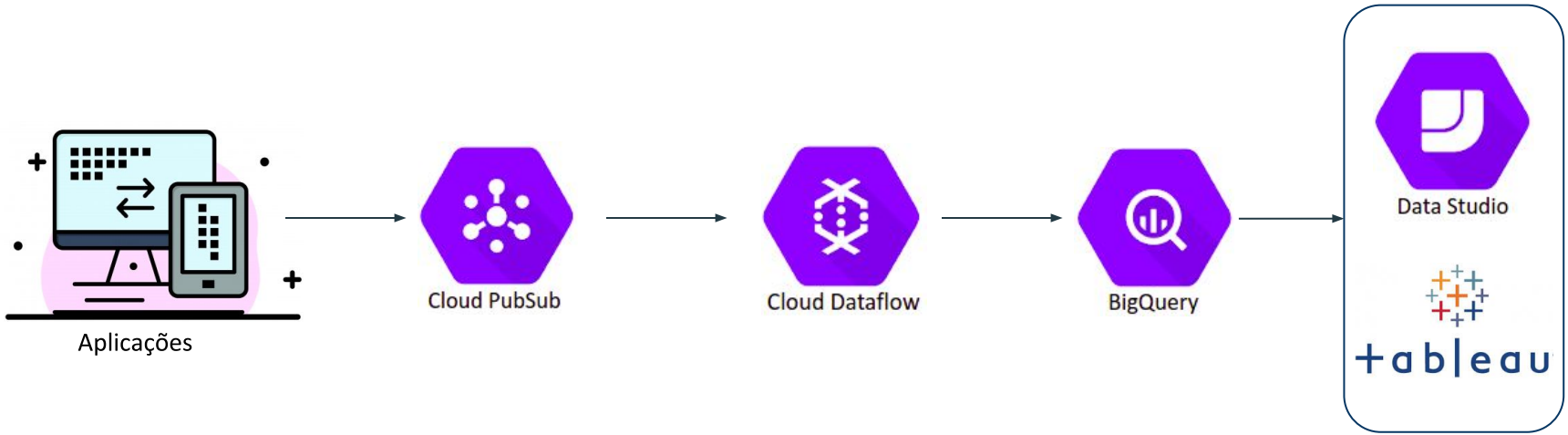
Ingestão pelo Witcher

O Witcher é um produto desenvolvido internamente em Pyspark para facilitar e automatizar a ingestão de dados no datalake. Automatiza a criação de dags e ajuda na diminuição de custo de cloud. Essa arquitetura utiliza streaming e batch.



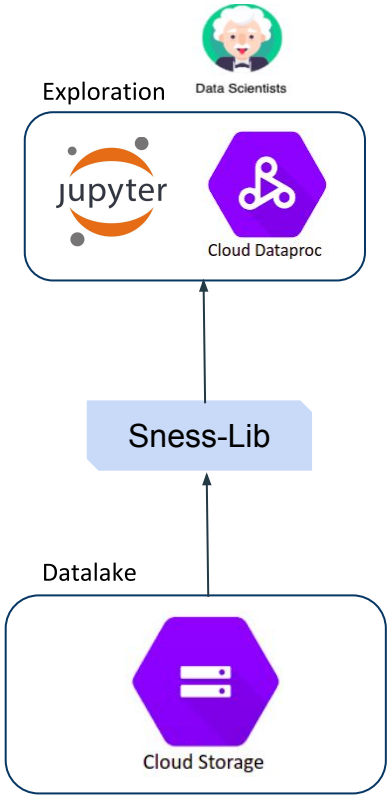
Ingestão pelo Niagara

O Niagara foi desenvolvido para aplicações que precisam do dados em tempo real para tomadas de decisões. Trabalha com recursos de Streaming utilizando mensageria Pubsub, e com processamento Dataflow (Serviço Apache Beam do Google) .



SNESS-LIB

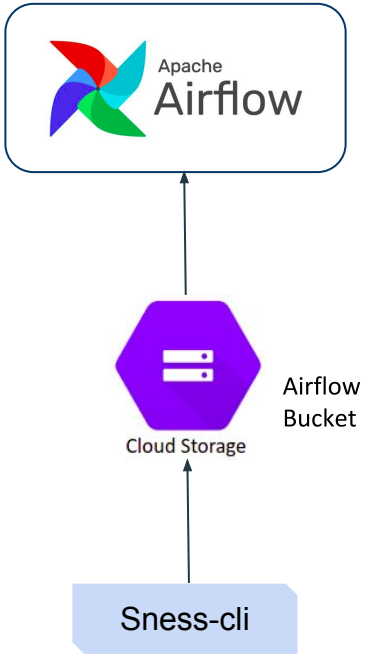
É uma Lib desenvolvida em Python para padronizar a leitura e escrita no Datalake. Além disso, possui alguns operadores customizados para Airflow que gerenciam a criação e deleção de clusters



luizalabs
MAGALU

SNESS-CLI

É um utilitário de linha de comando desenvolvido em python, para facilitar o deploy das Dags em produção. Também auxilia na criação de jobs sqoop.



luizalabs
MAGALU

ARCADE

BOT para gerenciamento de clusters de exploração sob demanda.

- Interface pelo Slack.
- Os clusters possuem integração com o Datalake
- Exploração de dados com Pyspark

luizalabs
MAGALU

DATA DIX

É um catálogo de dados da qual auxilia na governança dos dados. Possui metadados dos dataset existentes no datalake disponíveis ao usuários da plataforma e uma interface web para consulta dos mesmos.

Números da Plataforma

- Por volta de 40k execuções de DAGs por mês
- Por volta de 400 DAGs ativas no Airflow
- Cerca de 350 usuários, sendo que 180 são de Spark
- Mais de 4000 datasets
- Média de 15 Pb escaneados no BigQuery por mês
- Por volta de 700 Tb de dados no Lake