



CLOUDERA

Data Science bevezetés

Purger Norbert
Product Owner @ Cloudera

Budapest Data Forum 2020

ELJUTNI A FELFEDEZÉSTŐL A GYAKORLATI HASZNÁLATIG

Termékesített gépi tanulás, iparosodó mesterséges intelligencia



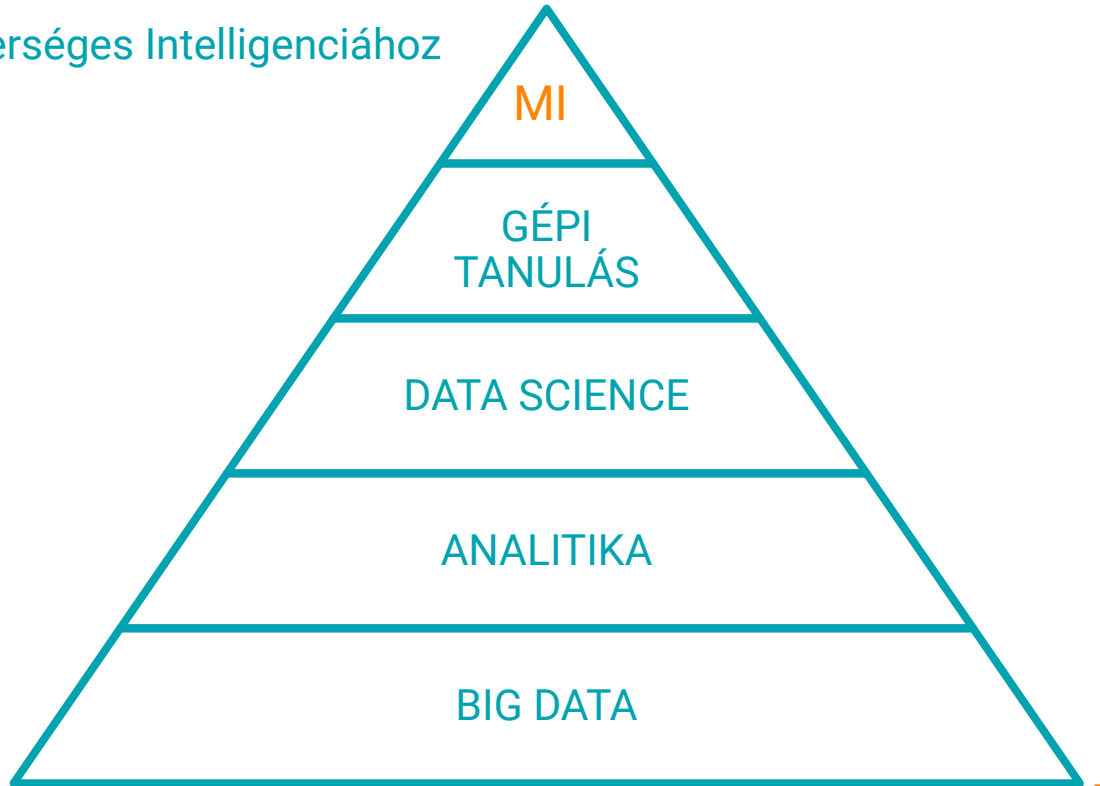
A kísérletektől ...



a gyárig.

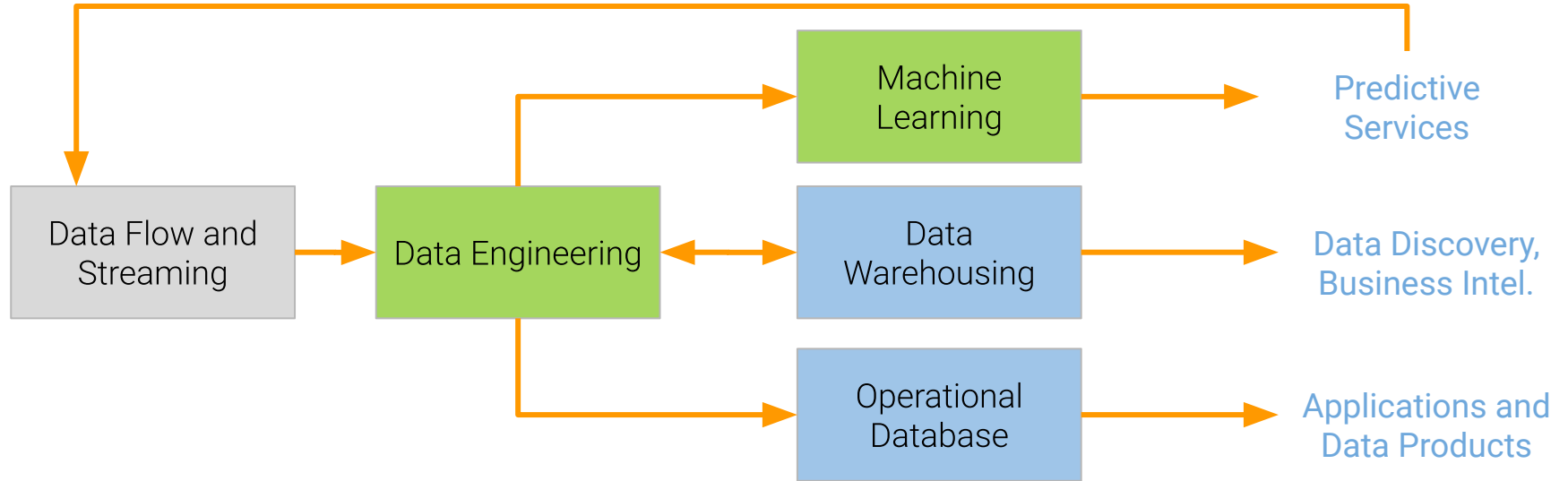
AZ ADATVEZÉRELT VÁLLALAT SZÜKSÉGLETHIERARCHIÁJA

Létra a Mesterséges Intelligenciához



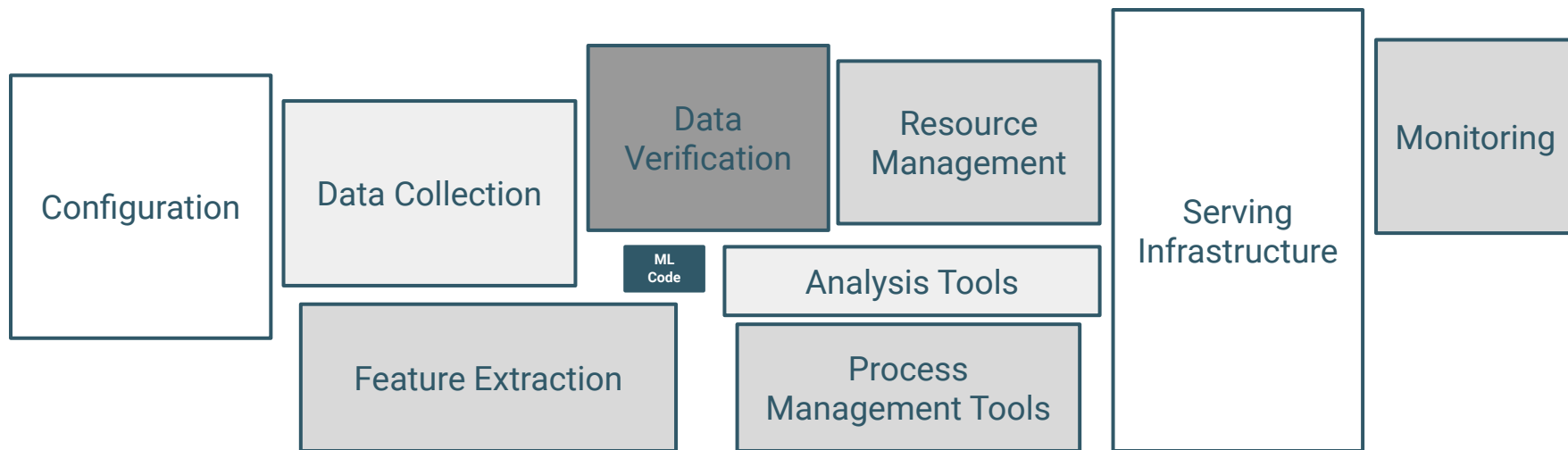
FOUNDATION OF MODERN DATA MANAGEMENT

CDP A Platform for Multiple Workloads



Data, Metadata, Security, Workload Management, Governance

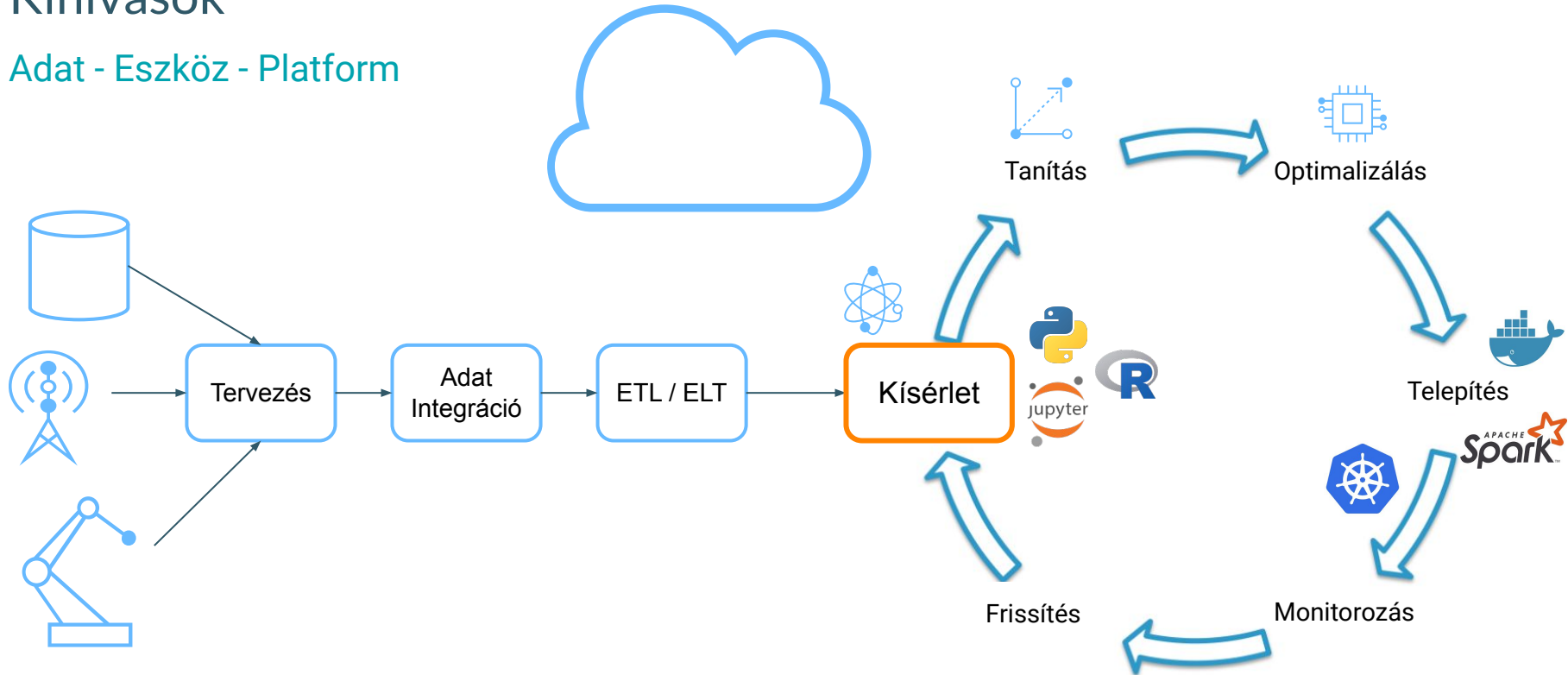
REJTETT TECHNIKAI ADÓSSÁG A GÉPI TANULÁSOS RENDSZEREKBEN



Forrás: <https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf>

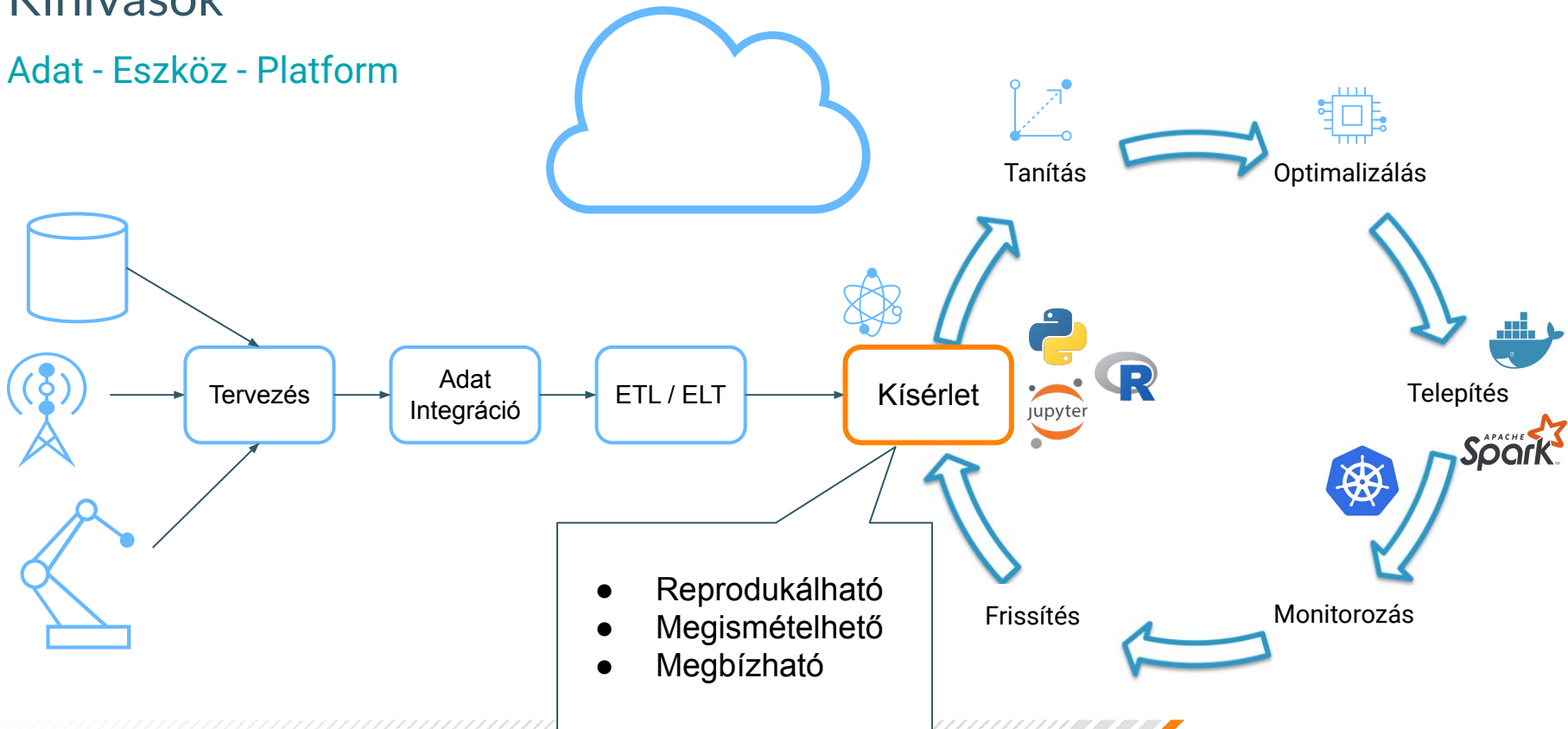
Kihívások

Adat - Eszköz - Platform



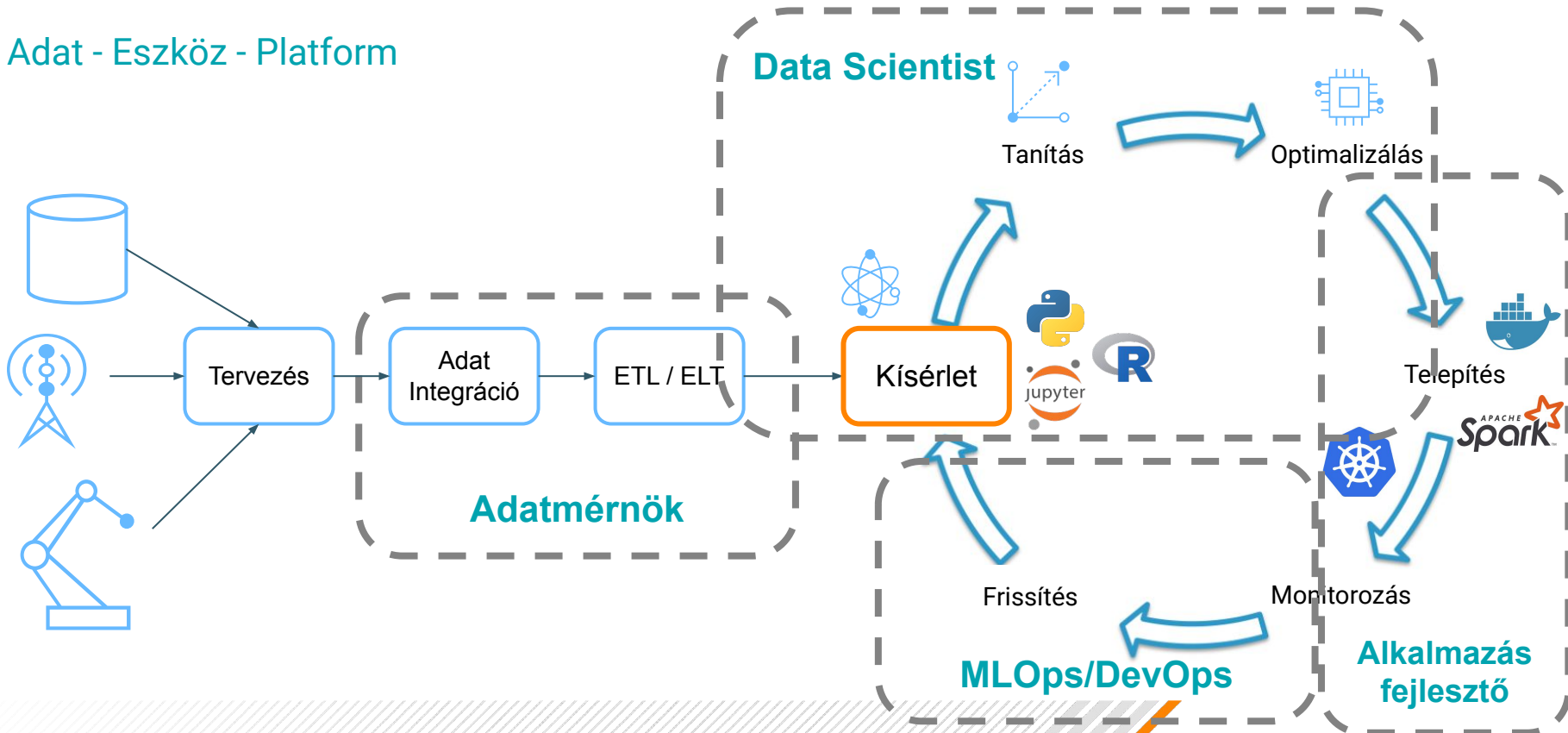
Kihívások

Adat - Eszköz - Platform



Kihívások

Adat - Eszköz - Platform



CSAPATJÁTÉK - SPECIALIZÁCIÓ ÉS IGÉNYEK



Adatmérnök

Kapacitás garancia,
konzisztens SLA



Üzleti elemző

Vizualizáció, igény
időszakos számítási
kapacításokra



Data Scientist

Legújabb szoftver, hardver
és kiszámíthatatlan
kapacítások



Alkalmazás fejlesztő

Folyamatos rendelkezésre
állítás, egyedi eszközök



Adat ~~É~~árépítész

Ő minden adatot ismer.



Infrastruktúra felelős

Megbízhatóság, költség,
skalázódás



Compliance Manager

Adatvédelem, Audit, GDPR

HASZNOS VÁLLALATI GÉPI TANULÁS

Csökkentsük a korlátokat a kísérletezéshez, a gépi tanulás építéséhez és telepítéséhez

1

Önkiszolgáló

Adjunk könnyű hozzáférést a data scientist csapatoknak a változatos adathoz, a szükséges számítási kapacitásokhoz és nyílt forrású eszközökhöz.

2

Munkafolyamat

Tegyük lehetővé a modern kollaboratív és reprodukálható szoftverfejlesztési gyakorlatokat és biztosítsunk több utat a termékesítéshez.

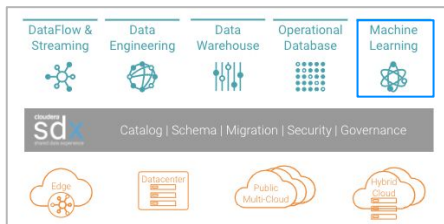
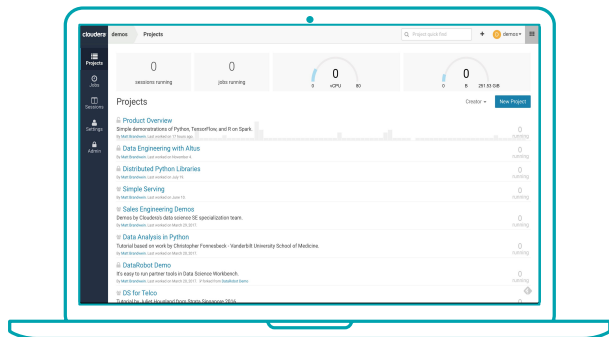
3

Bárhol

Ugyanolyan képességekkel a publikus és privát felhőben vagy a tradicionális adatkörnyezetekben.

CLOUDERA MACHINE LEARNING

Gyorsított gépi tanulás a kutatástól a termékesítésig



Adattudósoknak

Gyorsabb kísérletek

R, Python, Spark vagy Scala használat igény szerinti kapacitással és biztonságos CDH/HDP hozzáféréssel

Együttműködés

Reprodukálható eredmények megosztása az egész csapattal

Magabiztos start

Konzisztens lépések a termékesítésig újrakódolások nélkül

IT szakembereknek

Vigyünk az adattudományt az adathoz!

Nagyobb szabadság a kutató csapatnak, de mégis kisebb kockázat és kevesebb silósodás

Alapból biztonságos

Közös biztonság és adatkezelési elvek a különböző feladatok felett

Fusson bárhol

Helyben vagy a felhőben

WORLD'S FIRST ENTERPRISE DATA CLOUD: CLOUDERA DATA PLATFORM (CDP)

HYBRID & MULTI-CLOUD



DATA CENTER & PRIVATE CLOUD



HYBRID CLOUD



MULTI PUBLIC CLOUD

SECURITY & GOVERNANCE

CLUSTERA
SDX

METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE

ANALYTICS
EDGE TO AI



DATA HUB



DATA FLOW & STREAMING



DATA ENGINEERING



DATA WAREHOUSE



OPERATIONAL DATABASE



MACHINE LEARNING

OPEN
DISTRIBUTION

CLUSTERA RUNTIME



CONTROL PLANE



DATA CATALOG



REPLICATION MANAGER



WORKLOAD MANAGER



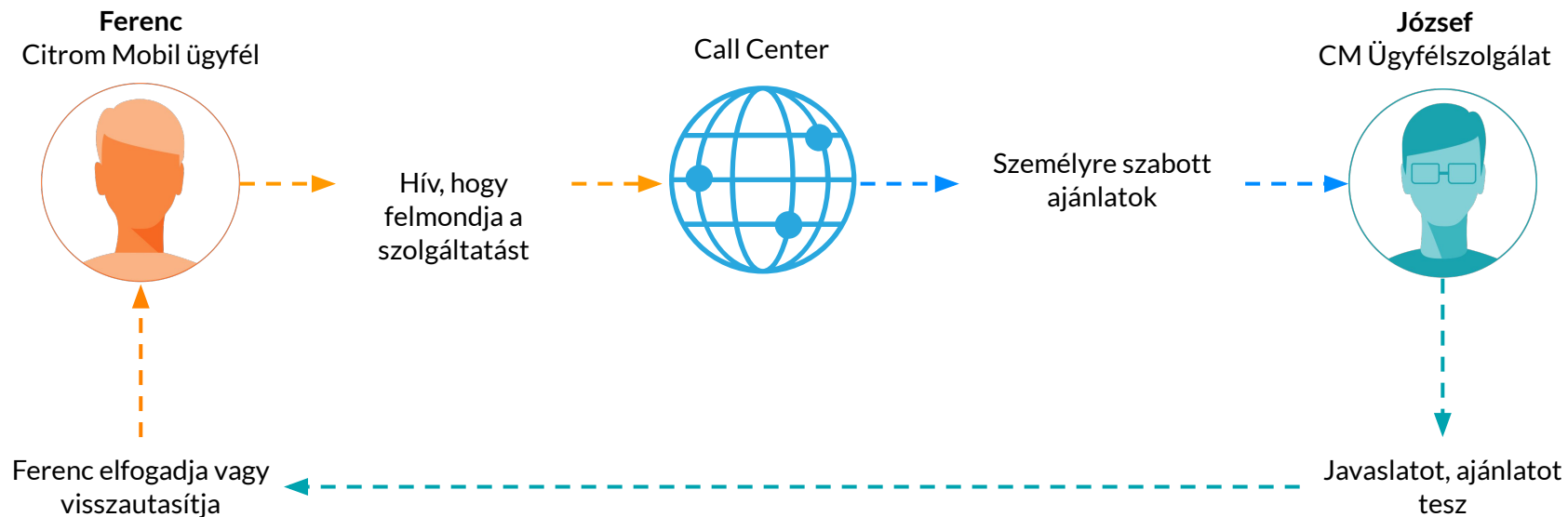
MANAGEMENT CONSOLE

DEMO



ÜGYFÉLSZOLGÁLAT

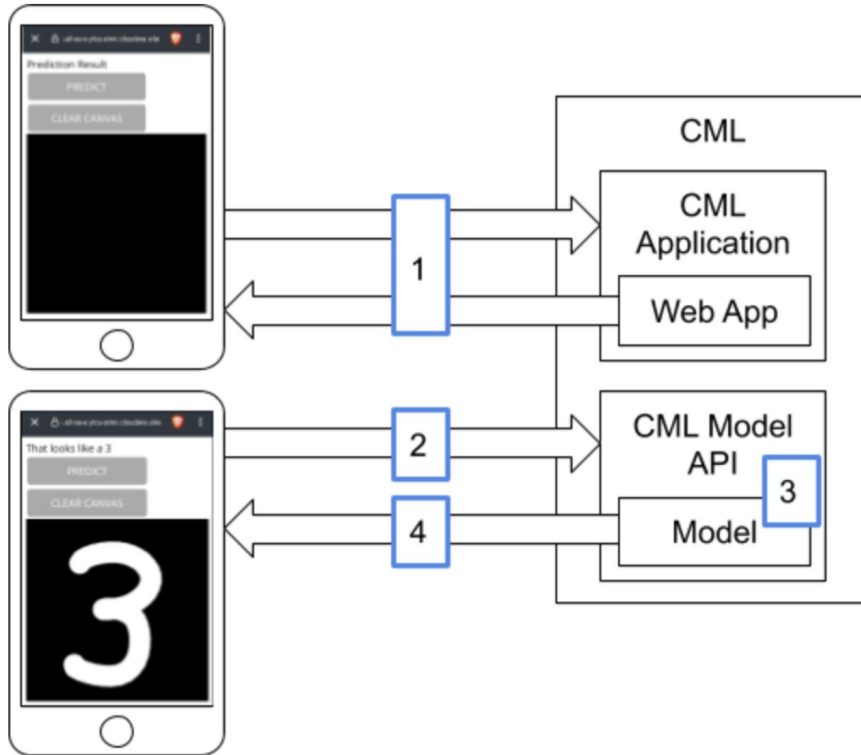
Ügyfélvesztés (churn) csökkentése model alapú predikcióval a call center folyamatokban



Alkalmazások



App + Model nagyon sok felhasználást tesz lehetővé



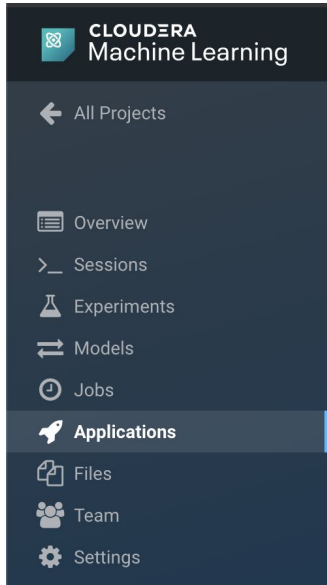
Teljes példa és blog [itt](#)

Alkalmazásfejlesztés választható keretrendszerben



Hosszan futó alkalmazások futtathatóak a project-en belül

st / Redirect App / Applications / New Application



Create an Application

General

Name *

Subdomain *

Description

Script *

A CML a megadott, jól ismert URL-en nyújthatja az alkalmazásokat

Rugalmasan futtatható web alkalmazások

i Choose a script that hosts a web application on either CDSW_READONLY_PORT or CDSW_APP_PORT. Both ports acts the same as CDSW_READONLY_PORT in a Session because in an Application container, CDSW_APP_PORT is the same as CDSW_READONLY_PORT. This means that a user with at least read access to the project can view the web application. However, please note that it is upto you to make the application really read-only in terms of entities belonging to the project. In other words, CDSW does not prevent you from running an application that allows a viewer to modify files that belongs to the project.

FEJLESZTÉS A VÁLASZTOTT KEDVENC FEJLESZTŐI KÖRNYEZETBEN

- Beépített, notebook jellegű szerkesztő
- Jupyter, mint teljes értékű környezet a klaszteren
- RStudio támogatás
- PyCharm elérés távoli/SSH kapcsolaton keresztül
- Apache Zeppelin támogatás hamarosan kipróbálható

Data scientist produktivitás

CML alapjaiban nyílt és bővíthető

auto-sklearn



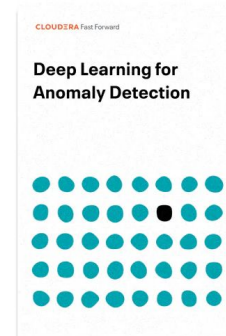
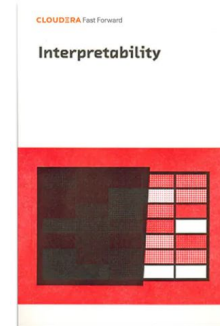
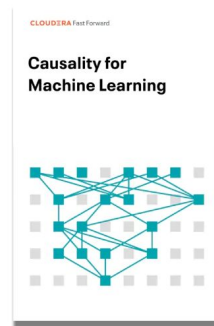
TransmogrifAI

automl-gs



HIVATKOZÁSOK

- CML & Cloudera Data Scientist Workbench (CDSW)
 - [CDP Public Cloud 60 napos próba](#)
 - [CDSW 1.7.2 \(on-prem\)](#)
- CML & Fast Forward Labs
 - GitHub - <https://github.com/fastforwardlabs>
 - [Research Reports](#)
 - Webinar & Blog
 - [Interpretálható ML Blog](#)
- Gyakorlat, gyakorlat, gyakorlat
 - Fast.AI
 - Kaggle



CML a hírekben

Cloudera Named Leader in The Forrester Wave:

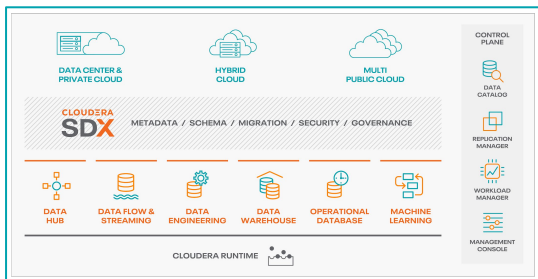
Notebook-Based Predictive Analytics and Machine Learning, Q3 2020

Cloudera is a machine learning - machine. Machine learning is the engine, but data...sweet, sweet data is the fuel

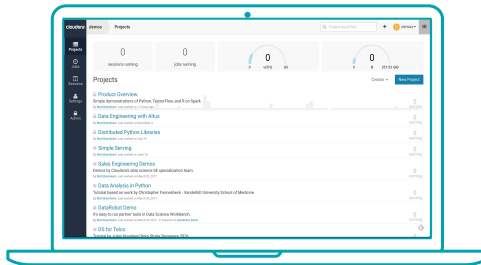


DATA SCIENCE & GÉPI TANULÁS A CLOUDERÁNÁL

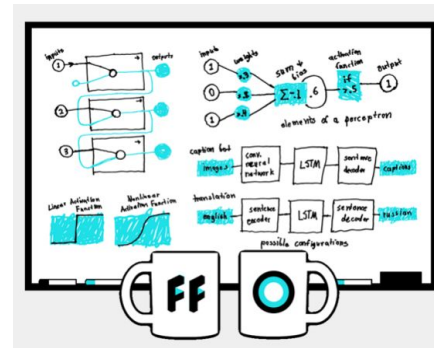
Hasznossá tesszük a vállalati gépi tanulást



Nyílt **platform** építeni, tanítani és működtetni a skálázódó gépi tanulási alkalmazásokat



Átfogó data science **eszköztárral** fokozni a csapat produktivitását



Szakértői támogatás és szolgáltatások a gyorsabb és szélesebb körű üzleti alkalmazásokért

THANK YOU

Q&A

CLOUDERA