بسم الله الرحمن الرحیم

**مستند سازی پروژه MLOps**

**شرکت آدین**

****

**اسفند ۱۴۰۱**

**نسخه ۰.۶**

فهرست

[۱. مشخصات کلی............................................................................................................................4](#_Toc127901742)

[نکات مثبت MLOP 5](#_Toc127901743)

[معایب MLOps 5](#_Toc127901744)

[Kubeflow چیست؟ 5](#_Toc127901745)

[معماری سرویس Kubeflow 6](#_Toc127901746)

[ویژگی های سرویس Kubeflow 6](#_Toc127901747)

[مقایسه Kubeflow و MLflow 7](#_Toc127901748)

[نمونه های جهانی استفاده از Kubeflow 7](#_Toc127901749)

[۲. معماری و تکنولوژی‌های مرتبط](#_Toc127901750) ...........................................................................................................8

[معماری مفهومی 8](#_Toc127901751)

[مرحله آماده سازی داده‌ها در چرخه حیات ML 9](#_Toc127901752)

[معماری منطقی 9](#_Toc127901753)

[معماری فیزیکی 10](#_Toc127901754)

[بهترین ابزار منبع باز برای جمع‌آوری داده ها 10](#_Toc127901755)

[ذخیره سازی داده‌ها 11](#_Toc127901756)

[بهترین دریاچه داده برای MLOps 14](#_Toc127901757)

[3. ویژگی‌های کارکردی](#_Toc127901758)

[شکاف‌های ارتباطی تیم‌های تجاری و فنی 16](#_Toc127901759)

[ارزیابی ریسک 16](#_Toc127901760)

[بازتاب تغییر اهداف کسب و کار در مدل 16](#_Toc127901761)

[۴. ویژگی‌های غیرکارکرد16.……………………………………………………………………………………………………………………](#_Toc127901762)

[آمادگی hybrid-cloud 16](#_Toc127901763)

[امنیت 16](#_Toc127901764)

[مقیاس پذیری 16](#_Toc127901765)

[تحمل خطا 16](#_Toc127901766)

[ارائه توان عملیاتی بالا 16](#_Toc127901767)

[پاسخ سریع 16](#_Toc127901768)

[۵. فرضیات و محدودیت‌ها ...................................................................................................................16](#_Toc127901769)

[۶. تست و ارزیابی MLOPs..................................................................................................................17](#_Toc127901770)

[تست مدل 17](#_Toc127901771)

[تست برنامه 17](#_Toc127901772)

[۷. آموزش و تحویل نصب و راه اندازی MLOPs............................................................................................17](#_Toc127901773)

[آموزش نصب Kubeflow 18](#_Toc127901774)

[راه اندازی Kubeflow 18](#_Toc127901775)

[داشبورد Kubeflow 22](#_Toc127901776)

[بخش های مختلف kubeflow به همراه نحوه کار آنها 22](#_Toc127901777)

[خط لوله kubeflow 25](#_Toc127901778)

[ایجاد و بارگذاری خط لوله 29](#_Toc127901779)

[تجربه اجرای یک خط لوله (Experiments kfp) 32](#_Toc127901780)

[Endpoint 33](#_Toc127901781)

[Kserve 33](#_Toc127901782)

[۸. منابع موردنیاز......................................................................................................................36](#_Toc127901783)

[۹. زمان‌بندی و شکست کارها........................................................................................................36](#_Toc127901784)

[۱۰. تحلیل هزینه.....................................................................................................................37](#_Toc127901785)

# مشخصات کلی

مدیریت سیستم‌های هوش مصنوعی در مقیاس بزرگ کار آسانی نیست و در این مسیر چالش‌های مهمی وجود دارد که تیم‌ها باید با آن مواجه شوند. برخی از این چالش‌ها در کاربرد یادگیری ماشین (ML) در یک محصول مقیاس‌پذیر عبارتند از:

* کیفیت و کمیت داده: اطمینان از وجود داده‌های آموزشی باکیفیت بالا برای آموزش مدل‌های ML یک چالش بزرگ است.
* عملکرد مدل: مدل‌های ML گاهی اوقات ممکن است برای تعمیم به داده‌های جدید مشکل داشته باشند و ممکن است در صورت استقرار در تنظیمات دنیای واقعی ضعیف عمل کنند.
* تفسیرپذیری مدل: مدل‌های ML می‌توانند پیچیده و غیرقابل‌درک باشند، و تفسیر پیش‌بینی‌ها و تشخیص مشکلات، آن‌ها را به چالش می‌کشد.
* منابع محاسباتی: آموزش و به‌کارگیری مدل‌های ML بزرگ و پیچیده می‌تواند به مقدار قابل‌توجهی از منابع محاسباتی نیاز داشته باشد که می‌تواند در یک محصول مقیاس‌پذیر چالش‌برانگیز باشد.
* استقرار مدل: استقرار مدل‌های ML در یک محیط تولید و ادغام آن‌ها در یک محصول می‌تواند چالش‌برانگیز باشد و نیاز به بررسی دقیق زیرساخت‌ها، حریم خصوصی داده‌ها و امنیت دارد.
* نگهداری مدل: مدل‌های ML باید با تغییر داده‌ها و نیازمندی‌های محصول، به‌روزرسانی و بازآموزی شوند که می‌تواند در یک محصول مقیاس‌پذیر یک چالش باشد.
* بایاس: مدل‌های ML گاهی اوقات می‌توانند سوگیری‌های موجود را در داده‌های آموزشی را تداوم بخشند و توجه به این مسائل هنگام توسعه یک محصول مقیاس‌پذیر ML بسیار مهم خواهد بود.

MLOps (عملیات یادگیری ماشین) به مجموعه‌ای از فرایندها، ابزارها و شیوه‌ها جهت مدیریت چرخه مدل‌های یادگیری ماشین در یک محیط تولید، اشاره دارد. همچنین این چرخه شامل همکاری بین دانشمندان داده، مهندسان و تیم‌های DevOps است به‌نحوی‌که این اطمینان حاصل شود که مدل‌ها به طور مؤثر توسعه، استقرار، نظارت و به‌روزرسانی می‌شوند. هدف MLOها افزایش سرعت، قابلیت اطمینان و مقیاس‌پذیری مدل‌های یادگیری ماشین در تولید است، درحالی‌که خطرات ناشی از ریسک عدم موفقیت را نیز کاهش می‌دهد. همچنین به‌کارگیری MLOps فرایند مدیریت را ساده‌تر کرده، کیفیت را افزایش می‌دهد و استقرار مدل‌های یادگیری عمیق و یادگیری ماشین را در محیط‌های تولید در مقیاس بزرگ را خودکار می‌کند. هدف، بهبود خودکارسازی و ارتقای کیفیت مدل‌های تولید و درعین‌حال توجه به الزامات تجاری و نظارتی است.

MLOps به‌ویژه در مورد استقرار مدل‌های ML در تولید اهمیت دارد، زیرا به سازمان‌ها کمک می‌کند تا مطمئن شوند که مدل‌هایشان در طول زمان دقیق، قابل‌اعتماد و کارآمد هستند. به‌طورکلی، MLOps با خودکار کردن بسیاری از مراحل مربوط به استقرار و مدیریت مدل‌های ML، به دانشمندان و مهندسان داده اجازه می‌دهد تا با همکاری یکدیگر به ارائه سریع‌تر و کارآمدتر مدل‌های یادگیری ماشین دست یابند. برای رسیدن به اهداف فوق، تیم‌های MLOps معمولاً از ترکیبی از ابزارها، فرایندها و شیوه‌ها زیر را استفاده می‌کنند:

* کنترل نسخه برای مدل‌های ML و مجموعه‌داده‌ها
* آزمون خودکار و ادغام مداوم
* کانتینرسازی و ارکستراسیون برای استقرار مدل
* نظارت و ثبت گزارش برای ردیابی عملکرد مدل در تولید
* حاکمیت و انطباق برای مدل‌ها

## MLOPsبرای چه مشتریانی مناسب است؟

* بهبود سرعت ایجاد و استقرار مدل در نتیجه مدیریت جامع چرخه زندگی یادگیری ماشین، به دلیل افزایش ارزش یادگیری ماشین.
* MLOها همکاری و شفافیت را در میان تیم‌های dataOps، مهندسین یادگیری ماشین، تحلیلگران تجاری/مدیران محصول، مهندسین تضمین کیفیت و مهندسین زیرساخت افزایش می‌دهند.
* MLOps زمان بازاریابی جهت الگوریتم‌های یادگیری ماشین را کاهش می‌دهد و روش‌های یکپارچه‌سازی و تحویل مداوم، استقرار این سیستم‌ها را در تولید آسان‌تر می‌کند. جهت اطمینان از ارائه پیش‌بینی‌های باکیفیت بالا، سیستم MLOps باید قادر به اندازه‌گیری دریفت مدل[[1]](#footnote-1) باشد. این امر احتمال دیدگاه‌های نادرست در فرضیه های مدل را کاهش می‌دهد.
* بهبود کیفیت مدل: با خودکارسازی آموزش، اعتبارسنجی و استقرار مدل‌ها، احتمال خطای انسانی کاهش می‌یابد که منجر به عملکرد بهتر مدل می‌شود.
* زمان سریع‌تر جهت بازاریابی: MLOps دانشمندان و مهندسان داده را قادر می‌سازد تا سریع‌تر تکرار کنند و مدل‌ها را سریع‌تر به تولید بفرستند.
* همکاری بهتر: MLOps همکاری بین دانشمندان داده، توسعه‌دهندگان و تیم‌های عملیاتی را تشویق می‌کند و تیم‌های متقابل را قادر می‌سازد تا به طور مؤثرتری با هم کار کنند.
* مقیاس‌پذیری: خودکار کردن فرایند استقرار، مقیاس‌پذیری مدل‌ها را جهت رسیدگی به مقادیر بزرگ و افزایشی داده و جمعیت بیشتر کاربران آسان‌تر می‌کند.

## چالش‌های MLOps

* **پیچیدگی**: پیاده‌سازی MLOها می‌تواند پیچیده باشد و به مهارت‌های تخصصی و درک عمیق فرایندهای یادگیری ماشین و توسعه نرم‌افزار نیاز دارد. MLOps به ابزارها و مهارت‌های فراوانی نیاز دارد که به پیچیدگی آن می‌افزاید. جهت ساختن یک چارچوب موفق MLOps، به یک معمار باتجربه MLOps و مهندسان ML نیاز داریم که از قبل با اصول مفهوم DevOps آشنا هستند.
* **چالش‌های یکپارچه‌سازی**: ادغام جریان‌های کاری ML در فرایندهای توسعه نرم‌افزار موجود می‌تواند دشوار باشد و ممکن است نیاز به تغییرات قابل‌توجهی در جریان‌های کاری و ابزارهای موجود داشته باشد.
* **امنیت داده‌ها**: MLOها به مدیریت دقیق داده‌های حساس از جمله ذخیره‌سازی ایمن، کنترل دسترسی و نظارت بر استفاده از داده‌ها نیاز دارند.
* **محدودیت‌های منابع:** پیاده‌سازی MLOها نیازمند سرمایه‌گذاری قابل‌توجهی در فناوری، فرآیندها و پرسنل است که ممکن است جهت همه سازمان‌ها امکان‌پذیر نباشد. MLOps جهت استقرار مؤثر مدل ML در دنیای واقعی ضروری است، اما هزینه آن به دلیل پیچیدگی و منابع موردنیاز می‌تواند بسیار گران باشد. توانایی ایجاد و حفظ یک چرخه حیات MLOps به طور مؤثر به منابع کافی (افراد، زیرساخت‌ها و ابزارها) نیاز دارد. عامل اصلی تعیین‌کننده در استفاده یا عدم استفاده از MLO ها، بررسی کامل بازگشت سرمایه ابتکار یادگیری ماشین است.

## Kubeflow چیست؟

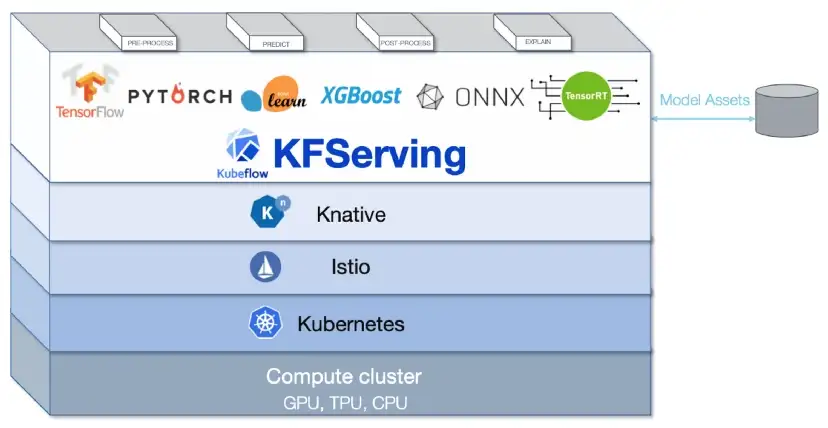
Kubeflow یک پلتفرم منبع باز برای توسعه، استقرار و مدیریت مدل‌های یادگیری ماشینی (ML) در Kubernetes (سیستم ارکستراسیون کانتینر) است به نحوی که توسعه، آزمایش و استقرار مدل‌های ML را در محیط‌های تولید برای سازمان‌ها آسان کرده و کارایی و مقیاس‌پذیری گردش‌های کاری ML خود را بهبود بخشند.

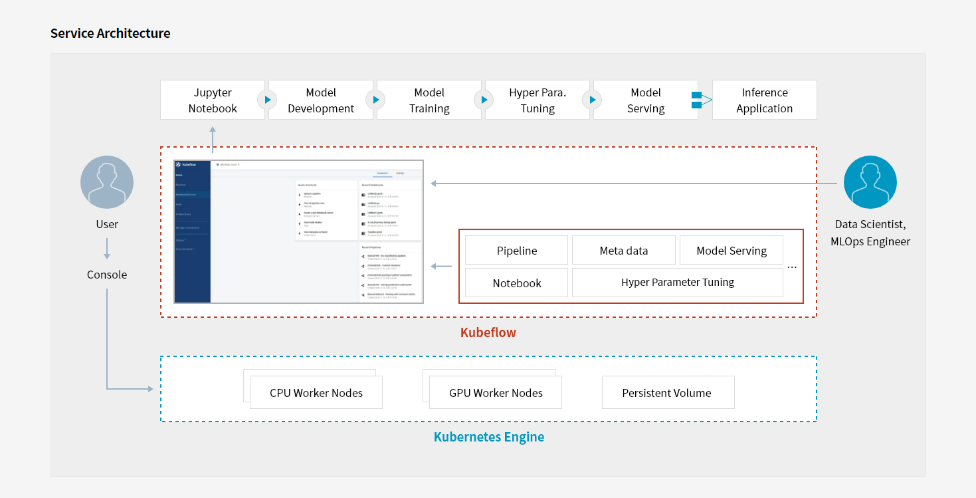
ویژگی‌های کلیدی Kubeflow عبارتند از:

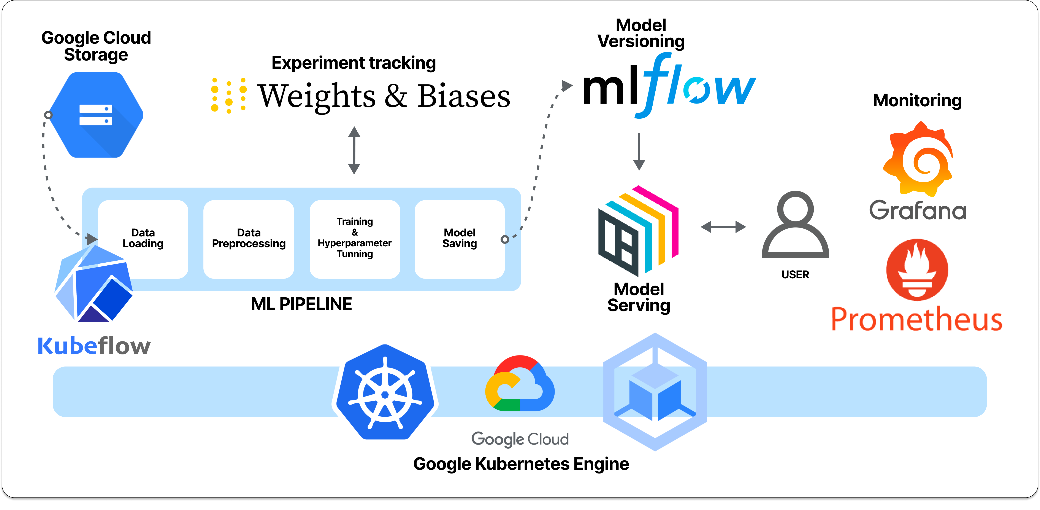
* یک رابط کاربر پسند برای آموزش و استقرار مدل‌های ML با استفاده از چارچوب‌های مطرح مانند TensorFlow، PyTorch و scikit-learn و ...
* پشتیبانی از آموزش توزیع‌شده مدل با استفاده از Kubernetes، که امکان آموزش سریع‌تر و استفاده بهتر از منابع را فراهم می‌کند.
* ابزارهایی برای ثبت و تجسم پیشرفت آموزش مدل، و همچنین برای تنظیم هایپرپارامتر و استقرار مدل.
* ادغام با کتابخانه‌ها و ابزارهای معروف یادگیری ماشین، مانند TensorFlow Extended (TFX) و Seldon و ...
* توانایی اجرای در سیستم شخصی، در فضای ابری یا در محیط های ترکیبی.
* به هدف استقرار کلیه کامپوننت های چرخه یادگیری ماشین بر بستر کوبرنتیز ساخته شده است. از این رو میتوان گفت یک مجموعه ابزار کاملا منطبق و همگام با زیست بوم ابری میباشد. همچنین برای CNCF و دریافت گواهینامه معتبر زیست بوم ابری این نهاد درخواست فرستاده است.

به‌طورکلی، Kubeflow یک پلت فرم قدرتمند و منعطف برای مدیریت چرخه یادگیری ماشین است که به دانشمندان و مهندسان داده اجازه می‌دهد تا با هم کار کنند تا مدل‌های یادگیری ماشینی خود را به شیوه‌ای کارآمدتر و مؤثرتر بسازند، مستقر کنند و مدیریت کنند.

## معماری سرویس Kubeflow







## ویژگی های سرویس Kubeflow

**محیط MLOps** **مبتنی بر ابر**

Kubeflow محیط‌های توسعه مدل ML را بهینه‌سازی شده برای ابر ارائه می‌دهد و امکان پیوند مبتنی بر Kubernetes را با نرم‌افزارهای منبع باز مختلف فراهم می‌کند.

**راحتی استفاده برای داده های بزرگ**

محیط های استاندارد از طیف وسیعی از چارچوب های یادگیری ماشین از TensorFlow، PyTorch، scikit-learn و Keras پشتیبانی می کنند. خط لوله برای کل فرایندهای توسعه، یادگیری و استقرار مدل‌های یادگیری ماشینی خودکار است تا از پیکربندی/ایجاد ساده و همچنین استفاده مجدد از مدل‌ها اطمینان حاصل شود.

**ویژگی های افزودنی**

Kubeflow منبع باز گسترده‌ای از اجرای کار یادگیری توزیعی/نظارت تا مدیریت خدمات استنتاج/تحلیل و مدیریت صف شغل در هر پلتفرمی ارائه می‌دهد. کاربران همچنین می‌توانند از زمان‌بندی‌های کار (FIFO، Bin-packing و Gang-based)، نظارت بر منابع GPU، گزارش موتور Kubeflow و ویژگی‌های اضافی دیگری که معمولاً در نرم‌افزارهای منبع باز در دسترس نیستند، استفاده کنند.

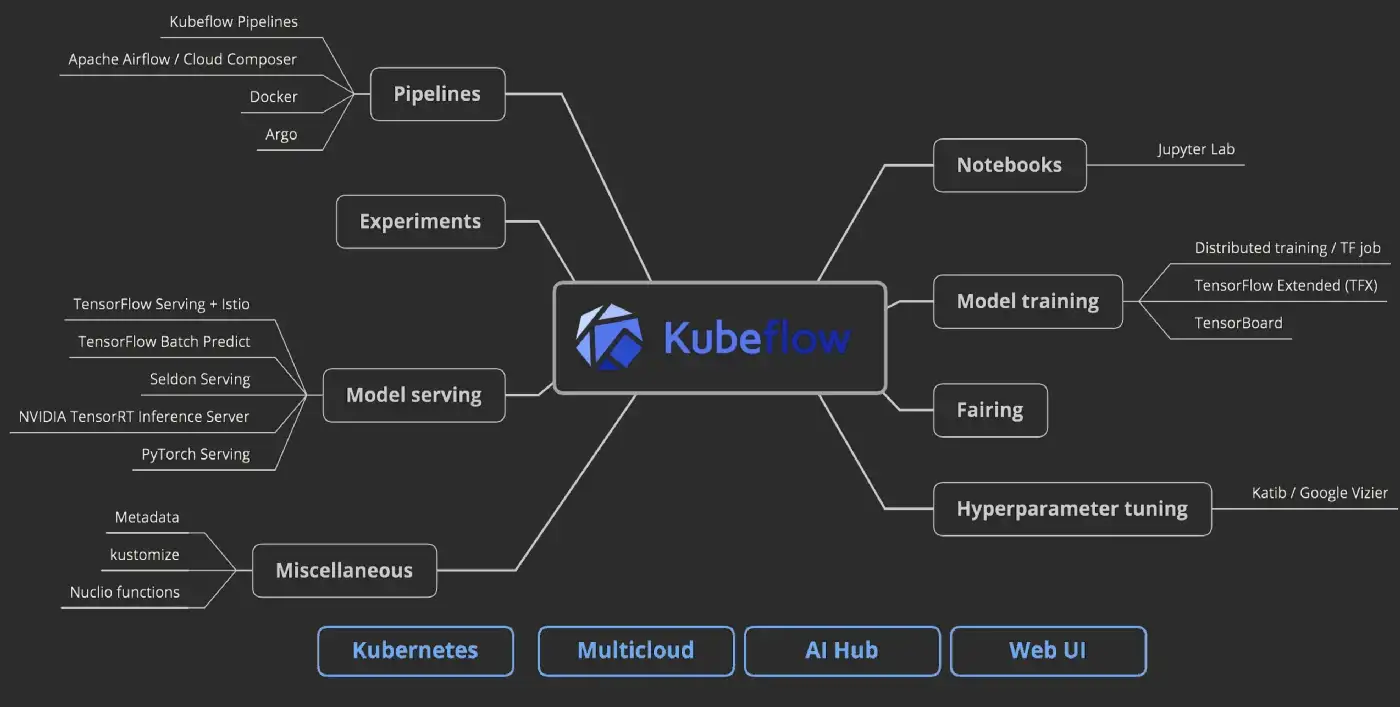
* سایر ویژگی ها
* نوت بوک Jupyter (توسعه مدل، یادگیری و استنتاج)
* اتوماسیون گردش کار (بر اساس خطوط لوله یادگیری ماشین)
* مدیریت زمانبندی کار GPU و صف کار
* نظارت بر منابع GPU
* مانیتورینگ/ ثبت نام Kubeflow و اجرای/نظارت کار یادگیری توزیع شده
* ساخت و مدیریت چارچوب ML (TensorFlow، PyTorch و غیره)
* مدیریت/تجزیه و تحلیل خدمات استنتاج و مدیریت آزمایش های مدل/گره های یادگیری

## چرا Kubeflow

Kubeflow و MLflow هر دو پروژه های منبع باز برای مدیریت گردش کار یادگیری ماشین هستند. با این حال، بررسی آنها در ابعاد مختلف نشان‌گر تفاوت‌های اساسی میان این دو پروژه است:

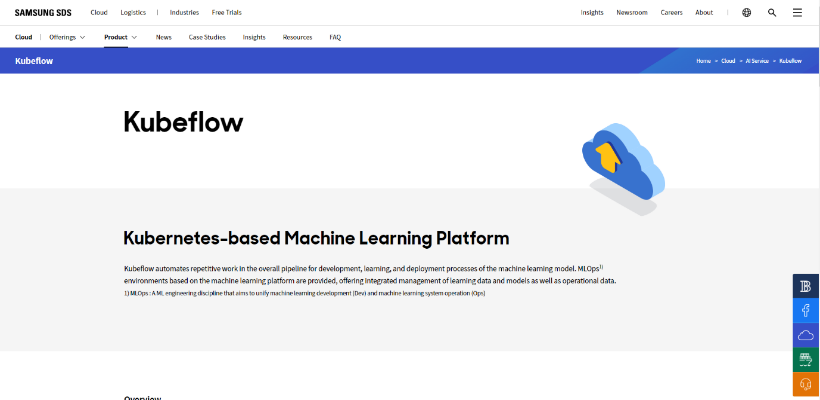
* هدف: Kubeflow به طور خاص برای اجرا و استقرار مدل‌های یادگیری ماشین در Kubernetes، یک سیستم منبع باز محبوب برای خودکارسازی استقرار، مقیاس‌بندی و مدیریت برنامه‌های کاربردی کانتینری طراحی شده است. از سوی دیگر، MLflow یک پلت فرم کلی تر برای مدیریت کل چرخه زندگی یادگیری ماشینی است، از جمله آزمایشات ردیابی، کد بسته بندی در اجراهای تکرارپذیر، و به اشتراک گذاری و استقرار مدل ها.
* ویژگی‌ها: Kubeflow مجموعه‌ای از ابزارها و APIها را برای ساخت، استقرار و مدیریت مدل‌های یادگیری ماشین در Kubernetes ارائه می‌کند، از جمله نوت‌بوک‌های Jupyter، آموزش TensorFlow و سرویس‌دهی مدل. MLflow طیف گسترده‌تری از ویژگی‌ها را برای چرخه زندگی یادگیری ماشین ارائه می‌کند، از جمله ردیابی آزمایش، بسته‌بندی کد و اشتراک‌گذاری، و استقرار مدل در پلتفرم‌های مختلف از جمله Docker، AzureML، و Amazon SageMaker.
* یکپارچه سازی: Kubeflow به خوبی با سایر اجزای اکوسیستم Kubernetes، مانند Helm (یک مدیر بسته برای Kubernetes) ادغام می شود، و همچنین ابزارهایی را برای کاهش حجم کاری یادگیری ماشین در مجموعه ای از ماشین ها فراهم می کند. MLflow را می توان با انواع کتابخانه ها و پلتفرم های یادگیری ماشین استفاده کرد و یک REST API برای ادغام با ابزارهای دیگر ارائه می دهد.

به طور خلاصه، Kubeflow برای مدیریت بارهای کاری یادگیری ماشین در Kubernetes مناسب است، در حالی که MLflow یک پلتفرم کلی تر برای مدیریت کل چرخه زندگی یادگیری ماشین است.



## نمونه های جهانی استفاده از Kubeflow

* از جمله شرکت ها و سازمان هایی که از Kubeflow به طور جدی برای آسان سازی فرانید چرخه یادگیری ماشین استفاده می‌کنند میتوان به مجموعه ای در شرکت سامسونگ با نام Samsung SDS اشاره کرد. در واقع این مجموعه Kubeflow را به عنوان پلتفرم ابری منطبق باKubernetes انتخاب کرده و به مشتریان خود بر بستر Samsung Cloud Platform پیاده‌سازی و خدمات ارائه میدهد.[[2]](#footnote-2)



* این مجموعه که زیر مجموعه ای از شرکت سامسونگ میباشد در زمینه‌های مختلفی سرویس ، آموزش و خدمات به Enterpriseها ارائه میکند. در زمینه‌ یادگیری ماشین به طور جدی از Kubeflow استفاده کرده و به صورت آموزش و در کنار شرکت های بزرگ همکاری میکند. در سایت این شرکت درباره سرویس AI خود به شرکت های بزرگ، Kubeflow را پلتفرم یادگیری ماشین مبتنی بر Kubernetes معرفی می کند.
* از جمله شرکت های بزرگ دیگر میتوان به Arrikto اشاره کرد. این شرکت مستقر در امریکا در واقع برای یادگیری ماشین Kubeflow as a Service ارائه میدهد.
* Arrikto، برخلاف سایر شرکت‌های MLOps، تیم‌های MLOps را قادر می‌سازد تا مدل‌های یادگیری ماشین را 30 برابر سریع‌تر از پلتفرم‌های سنتی ML به بازار عرضه کنند. Arrikto به عنوان یک مشارکت‌کننده پیشرو در Kubeflow، گردش‌های کاری خودکار، خطوط لوله قابل تکرار، استقرار مداوم از دسک‌تاپ به ابر و دسترسی ایمن به داده‌ها را ارائه می‌کند. Arrikto's Enterprise Kubeflow به عنوان یک توزیع چند گره در AWS، GCP و Azure در دسترس است و پلتفرم ترجیحی MLOps است که امروزه توسط بسیاری از شرکت های Fortune 500 در تولید استفاده می شود. این شرکت بیش از 300 مشتری در 17 کشور جهان دارد و توسط Unusual Ventures و Odyssey VP حمایت می شود.

# معماری و تکنولوژی‌های مرتبط

## معماری مفهومی

به‌طورکلی مراحل اصلی در معماری مفهومی MLOpsعبارت‌اند از:

1. **Data Preparation - آماده سازی داده‌ها**:

این مرحله شامل جمع‌آوری، تمیز کردن و پیش پردازش داده‌هایی است که برای آموزش و ارزیابی مدل‌های ML استفاده می‌شود. آماده‌سازی داده‌ها ممکن است شامل کارهایی مانند عادی‌سازی داده‌ها، تشخیص نقاط پرت و مهندسی ویژگی‌ها باشد.

1. **Model Selection - انتخاب مدل**:

در این مرحله، دانشمندان و مهندسان داده، معماری، الگوریتم و فراپارامترهای مدل ML مناسب را بر اساس ویژگی‌های داده‌ها و مسئله‌ای که باید حل شود، انتخاب می‌کنند.

**Model Training - آموزش مدل**:

این مرحله شامل استفاده از داده‌های آماده شده برای آموزش مدل انتخاب شده، با تنظیم پارامترهای مدل برای به حداقل رساندن خطا بین خروجی‌های پیش بینی شده و خروجی‌های واقعی است.

1. **Model Evaluation - ارزیابی مدل**:

این مرحله شامل ارزیابی عملکرد مدل آموزش‌دیده با مقایسه پیش‌بینی‌های آن با مجموعه‌ای از نتایج شناخته‌شده (که داده‌های آزمایشی نیز نامیده می‌شود)، و تعیین میزان تعمیم مدل به داده‌های نادیده جدید است.

1. **Model Deployment - استقرار مدل**:

این مرحله شامل در دسترس قرار دادن مدل برای کاربران نهایی است، خواه از طریق استقرار مدل در یک محیط تولید یا با ارائه یک API برای دسترسی دیگران به مدل باشد.

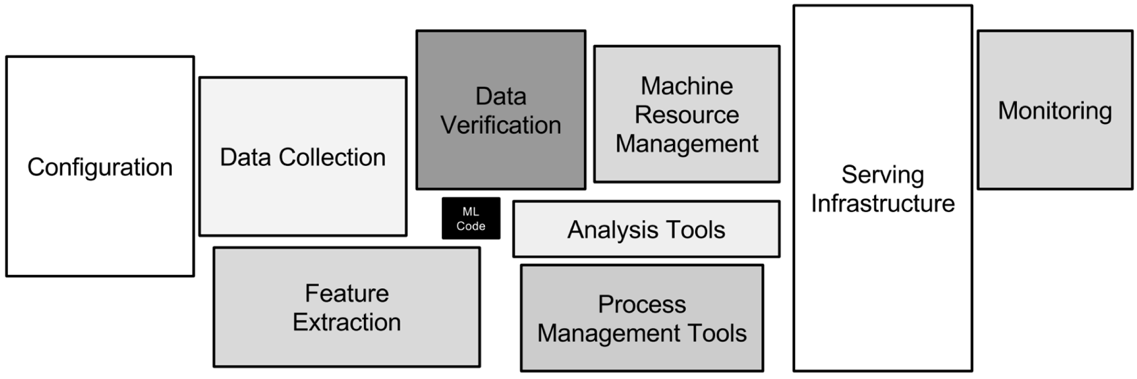
1. **Model Monitoring and Maintenance - نظارت و نگهداری مدل:**

این مرحله شامل نظارت بر عملکرد مدل مستقر در طول زمان و انجام به روز رسانی‌ها یا تنظیمات موردنیاز برای حفظ دقت و عملکرد مدل است.

**Model Retainment- بازنشستگی مدل:**

این مرحله شامل بازنشستگی مدل‌هایی است که دیگر موردنیاز نیستند یا دیگر عملکرد خوبی ندارند، تا اطمینان حاصل شود که فقط مدل‌هایی با عملکرد خوب به کار گرفته می‌شوند.

به‌طورکلی، هر یک از این مراحل نقش مهمی در چرخه زندگی یادگیری ماشین ایفا می‌کند و برای سازمان‌ها مهم است که یک فرآیند MLOps قوی و کارآمد را برای مدیریت چرخه عمر ML از ابتدا تا انتها اجرا کنند.

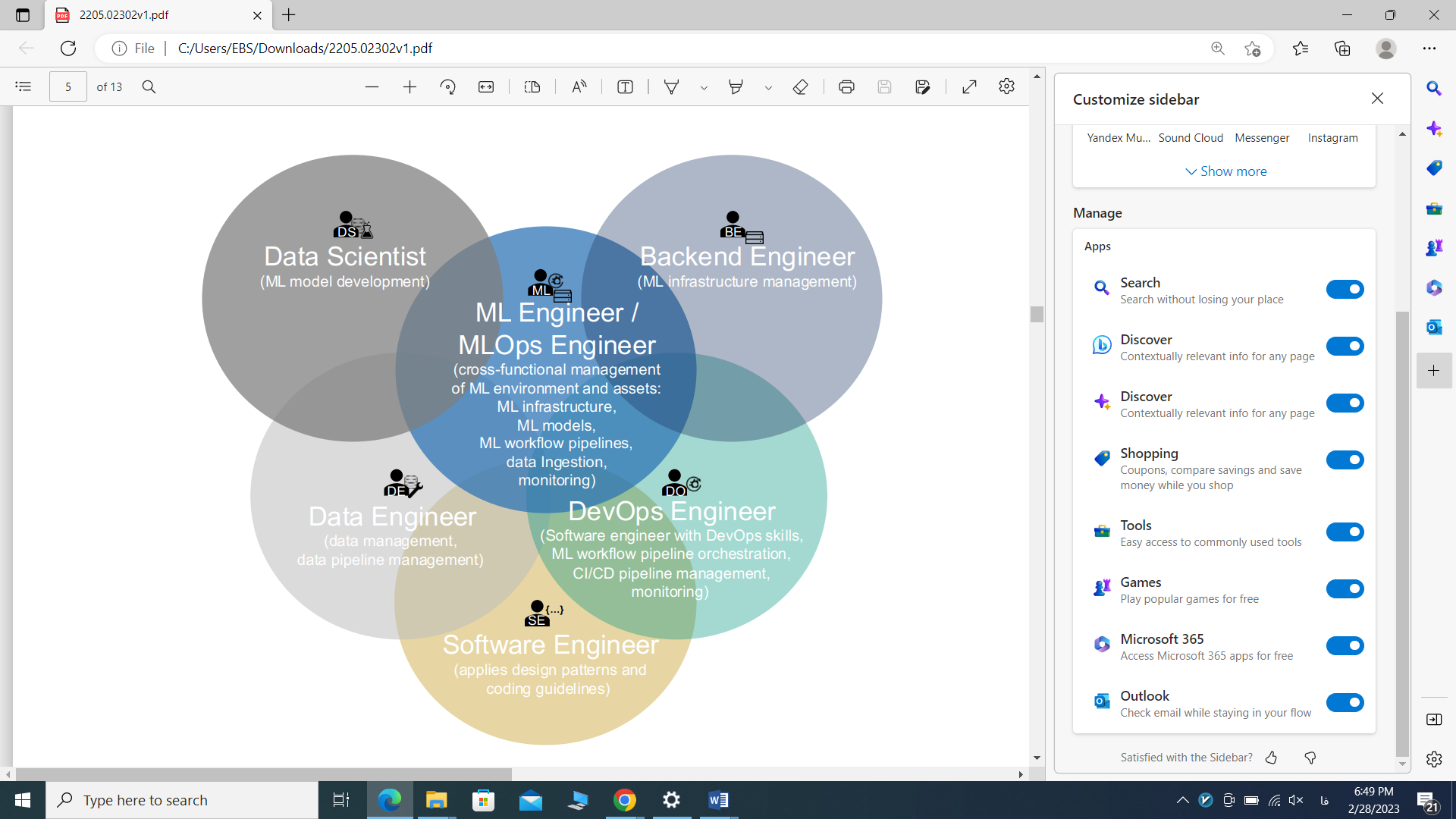


آماده سازی داده‌ها اولین مرحله در چرخه حیات یادگیری ماشینی (ML) است و شامل جمع‌آوری، تمیز کردن و پیش پردازش داده‌هایی است که برای آموزش و ارزیابی مدل‌های ML استفاده می‌شود. هدف از این مرحله آماده سازی داده‌ها در قالبی است که به راحتی توسط مدل ML مصرف شود و اطمینان حاصل شود که داده‌ها از کیفیت بالایی برخوردار هستند.

## مرحله آماده سازی داده‌ها در چرخه حیات ML

برخی از وظایف متداول درگیر در تهیه داده‌ها عبارتند از:

1. **Data Collection - جمع‌آوری داده‌ها**: جمع‌آوری داده‌ها از منابع مختلف مانند پایگاه‌های داده، فایل‌ها و APIهای خارجی.
2. **Data Cleaning - پاکسازی داده‌ها**: شناسایی و حذف هر گونه خطا، ناسازگاری یا موارد پرت در داده‌ها و اطمینان از کامل و دقیق بودن داده‌ها
3. **Data Integration - یکپارچه سازی داده‌ها:** ترکیب چندین منبع داده در یک مجموعه داده واحد، و حل هر گونه تضاد یا ناسازگاری که ممکن است ایجاد شود.
4. **Data Transformation - تبدیل داده‌ها:** تبدیل داده‌ها به قالبی که می‌تواند توسط مدل ML مصرف شود، مانند تبدیل متغیرهای طبقه بندی به مقادیر عددی.
5. **Data Normalization - عادی سازی داده‌ها**: مقیاس بندی داده‌ها به گونه ای که محدوده مقادیر یکسانی داشته باشند، که می‌تواند به بهبود عملکرد برخی از مدل‌های ML کمک کند.
6. **Feature Engineering - مهندسی ویژگی:** ایجاد ویژگی‌ها یا متغیرهای جدید که می‌تواند به بهبود عملکرد مدل ML کمک کند.
7. **Data Splitting - تقسیم داده‌ها:** تقسیم داده‌ها به مجموعه‌های آموزشی، اعتبار سنجی و تست، که برای آموزش، ارزیابی و آزمایش مدل ML استفاده می‌شود.

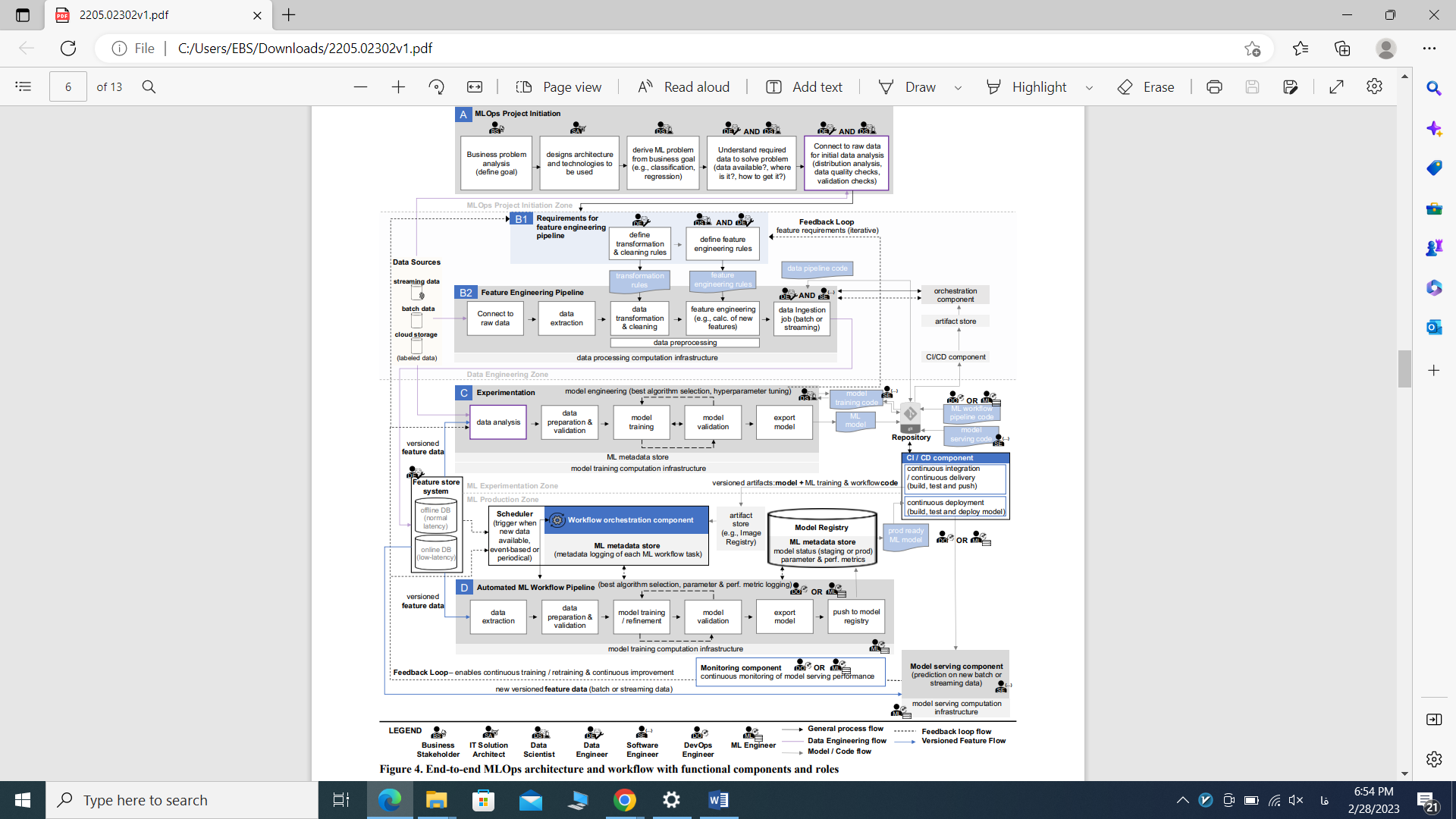


به‌طورکلی، آماده‌سازی داده‌ها مرحله‌ای حیاتی در چرخه حیات ML است، زیرا کیفیت و قالب داده‌ها به شدت بر دقت و عملکرد مدل نهایی ML تأثیر می‌گذارد. یک مجموعه داده به خوبی آماده شده می‌تواند به یک مدل دقیق و کارآمد منجر شود، و بنابراین، سرمایه‌گذاری زمان و منابع برای اطمینان از پاکسازی، تبدیل و آماده شدن داده‌ها برای مراحل بعدی چرخه حیات ML مهم است.

## معماری منطقی

معماری منطقی MLOps معمولاً شامل اجزای زیر است:

* مدیریت کد منبع: این مؤلفه کد منبع مدل‌های یادگیری ماشین، الگوریتم‌ها و خطوط لوله را ذخیره می‌کند.
* یکپارچه سازی مداوم/ استقرار مستمر (CI/CD): این جزء ادغام و استقرار مدل‌های یادگیری ماشین را خودکار می‌کند.
* آموزش مدل: این جزء با استفاده از داده‌ها و الگوریتم‌های ذخیره شده در جزء مدیریت کد منبع، مدل‌های یادگیری ماشین را آموزش می‌دهد.
* مدیریت مدل: این مؤلفه مدل‌های آموزش‌دیده و فراداده‌های مربوط به مدل‌ها مانند نسخه، دقت و داده‌های آموزشی را ذخیره می‌کند.
* استقرار مدل: این جزء، مدل‌های آموزش دیده را در یک محیط تولید مستقر می‌کند، جایی که می‌توان از آنها برای پیش بینی استفاده کرد.
* مانیتورینگ مدل: این مؤلفه عملکرد مدل‌های مستقر شده را نظارت می‌کند و در مورد دقت و قابلیت اطمینان آنها بازخورد ارائه می‌کند.
* مدیریت مدل: این مؤلفه چرخه عمر مدل‌های یادگیری ماشین را مدیریت می‌کند، از جمله نسخه‌سازی، به‌روزرسانی و بازنشستگی مدل‌ها در صورت لزوم.



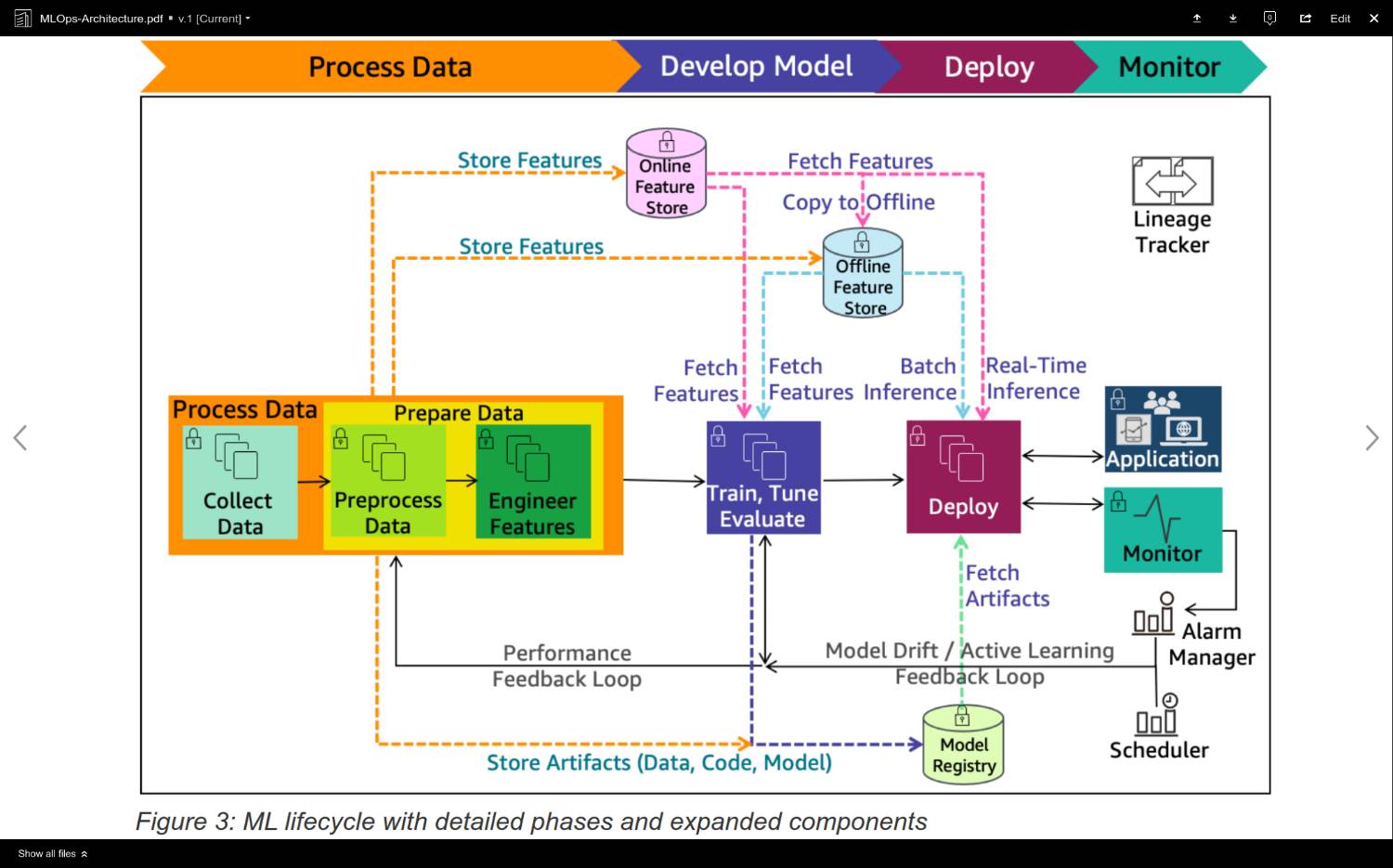
## معماری فیزیکی

معماری فیزیکی MLOها به نیازهای خاص سازمان و پروژه‌های یادگیری ماشینی در حال توسعه بستگی دارد. زیرساخت باید مقیاس پذیر، انعطاف پذیر و قادر به رسیدگی به نیازهای جریان کار یادگیری ماشین و ادغام و استقرار مداوم مدل‌های یادگیری ماشین باشد. معماری فیزیکی MLOps به زیرساخت فیزیکی، سخت افزار و اجزای نرم افزاری مورد استفاده برای پیاده‌سازی معماری منطقی MLOها اشاره دارد. این می‌تواند شامل موارد زیر باشد:

* زیرساخت محاسباتی: این می‌تواند شامل سرورهای داخلی، ماشین‌های مجازی مبتنی بر ابر یا ترکیبی از هر دو باشد.
* ذخیره‌سازی داده‌ها: این می‌تواند شامل پایگاه‌های داده، دریاچه‌های داده یا سیستم‌های فایل مورد استفاده برای ذخیره داده‌های آموزشی، داده‌های آزمایشی و خروجی‌های مدل باشد.
* شبکه سازی: این می‌تواند شامل اجزای سخت افزاری و نرم افزاری باشد که از ارتباط بین اجزای مختلف معماری MLOps پشتیبانی می‌کند.
* نظارت: این می‌تواند شامل ابزارهای نظارت، تحلیلگرهای گزارش و داشبوردهایی باشد که برای نظارت بر عملکرد مدل‌های یادگیری ماشین و گردش کار MLOps استفاده می‌شوند.
* امنیت: این می‌تواند شامل فایروال‌ها، سیستم‌های کنترل دسترسی، رمزگذاری و سایر اقدامات امنیتی برای محافظت از داده‌ها و مدل‌های حساس درگیر در MLOها باشد.
* ابزارها و پلتفرم‌ها: این می‌تواند شامل ابزارها و پلتفرم‌های نرم‌افزاری منبع باز و تجاری باشد که برای یادگیری ماشین استفاده می‌شوند، مانند TensorFlow، PyTorch و غیره.

برخی ابزار/پلتفرم برتر MLOps جهت مدیریت چرخه یادگیری ماشین عبارتنداز:

* Amazon SageMaker
* Azure Machine Learning
* TensorFlow Extended (TFX)
* MLFlow
* Google Cloud ML Engine
* Data Version Control (DVC)
* H2O Driverless AI
* Kubeflow
* و ...



# ویژگی‌های کارکردی

## شکاف‌های ارتباطی تیم‌های تجاری و فنی

چالش اصلی بین تیم‌های تجاری و فنی ارتباط است. پیدا کردن زبان مشترک جهت همکاری سخت است. بیشتر اوقات این شکاف دلیل شکست پروژه‌های بزرگ است.

## ارزیابی ریسک

ارزیابی ریسک نیز یکی از چالش‌های مهم است، زیرا مدل‌ها اغلب تمایل دارند از آنچه در ابتدا نتایج می‌دادند دور شوند. ارزیابی هزینه/خطر چنین شکست‌هایی یک کار ضروری و چالش برانگیز است.

## بازتاب تغییر اهداف کسب و کار در مدل

وابستگی‌های زیادی با داده‌ها وجود دارد که به طور مداوم عملکرد را حفظ می‌کند، استانداردهای مدل را تغییر می‌دهد و از حاکمیت هوش مصنوعی اطمینان می‌دهد. همگام شدن با اهداف در حال تحول کسب و کار و آموزش مدل مستمر چالش برانگیز است.

# ویژگی‌های غیرکارکردی

## آمادگی hybrid-cloud

## امنیت

## مقیاس پذیری

## تحمل خطا

## ارائه توان عملیاتی بالا

## پاسخ سریع

# فرضیات و محدودیت‌ها

MLOps جهت استقرار مؤثر مدل ML در دنیای واقعی ضروری است، اما هزینه آن به دلیل پیچیدگی و منابع موردنیاز می‌تواند بسیار گران باشد. عامل اصلی تعیین‌کننده در استفاده یا عدم استفاده از MLOps ها، بررسی کامل بازگشت سرمایه با بگارگیری یادگیری ماشین است.

* پیاده‌سازی MLOpsها ممکن است گران باشد.
* توانایی ایجاد و حفظ یک چرخه حیات MLOps به طور مؤثر به منابع کافی (افراد، زیرساخت‌ها و ابزار) نیاز دارد.
* MLOps اگر به درستی انجام نشود، به ابزارها و مهارت‌های فراوانی نیاز دارد که بر پیچیدگی آن افزوده می‌شود.
* جهت ساختن یک چارچوب موفق MLOps ، به یک معمار باتجربه MLOps و مهندسان ML نیاز داریم که از قبل با اصول مفهوم DevOps آشنا هستند.

# تست و ارزیابی MLOPs

پس از ایجاد تست‌ها، می‌توانیم هر زمان که تغییری در سیستم ایجاد شد، تست‌ها را به‌طور خودکار اجرا شود و در طول زمان به بهبود آنها ادامه دهیم. جهت جلوگیری از افزایش هزینه‌های پایین دستی و از دست دادن زمان، پاداش دادن به اجرای تست‌ها و شناسایی منابع اشتباه در سریع‌ترین زمان ممکن در چرخه توسعه، مناسب خواهد بود.

## ارزیابی Kubeflow Pipelines

## اسکریپت‌های ارزیابی Kubeflow Pipelines بار کاری معمولی را شبیه‌سازی می‌کنند و معیارهای عملکردی مانند تأخیرهای سرور و مدت زمان اجرای خط لوله را ثبت می‌کنند.

## برای شبیه‌سازی یک بار کاری معمولی، اسکریپت معیار یک فایل مانیفست خط لوله را در نمونه‌ای از Kubeflow Pipelines به عنوان خط لوله یا نسخه خط لوله آپلود می‌کند و چندین اجرا را به طور همزمان ایجاد می‌کند. می توان مانیفست خطوط لوله مورد استفاده در اسکریپت معیار را مشخص کرد، یا می توان از یکی از خطوط لوله نمونه از پیش بارگذاری شده در Kubeflow استفاده کرد.

## علاوه بر این، استفاده از یک مانیفست خط لوله که نماینده مورد استفاده خاص شما باشد، نیز تمرین خوبی است. به عنوان مثال، اگر خوشه خطوط لوله Kubeflow شما عمدتاً برای خطوط لوله وظایف تشخیص تصویر استفاده می شود، پس بهتر است از خط لوله تشخیص تصویر در اسکریپت های معیار استفاده شود.

## همانطور که قبلا ذکر شد، پس از انتخاب خط لوله مناسب، اسکریپت های Benchmark آن را چندین بار به طور همزمان اجرا می کنند. در میان تمام عملیاتی که خطوط لوله Kubeflow می توانند انجام دهند، اجرای یک خط لوله بدون شک غیرقابل پیش بینی ترین و پرهزینه ترین است.

## سایر عملیات، مانند ایجاد خط لوله (نسخه) یا ایجاد یک آزمایش، معمولاً هزینه قابل پیش بینی و متوسطی را القا می کند. به عنوان مثال، ایجاد یک نسخه خط لوله، یک ردیف جدید در جدول نسخه های خط لوله و یک فایل جدید در سرور MinIO معرفی می کند. اندازه فایل جدید به مانیفست نسخه خط لوله بستگی دارد. اگر مورد نادر یک مانیفست بسیار بزرگ را کنار بگذاریم و یک مانیفست با اندازه متوسط را برای هر نسخه خط لوله ایجاد شده فرض کنیم، هزینه کل ایجاد یک نسخه خط لوله به صورت خطی با تعداد نسخه های خط لوله افزایش می یابد. با این حال، از سوی دیگر، هزینه اجرای یک خط لوله یا نسخه خط لوله شامل عدم اطمینان بسیار بیشتر و گاهی اوقات هزینه بسیار بالایی است.

## یک خط لوله یا یک نسخه خط لوله می تواند دارای اجزای دلخواه باشد و از این رو اجرای یک خط لوله یا نسخه خط لوله می تواند پیچیدگی های زمانی و مکانی دلخواه را به همراه داشته باشد. به عنوان مثال، یک مرحله در یک خط لوله می تواند از یک تصویر ظرف سفارشی استفاده کند که یک کار آموزشی بسیار گران قیمت را انجام می دهد. علاوه بر این، اجراها در نمونه Kubeflow Pipelines نیز فضای DB بیشتری را نسبت به خطوط لوله، نسخه‌های خط لوله، آزمایش‌ها و غیره مصرف می‌کنند. بنابراین، به منظور درک عملکرد و مقیاس پذیری نقاط درد در نمونه Kubeflow Pipelines، تمرکز بر عملیات اجرا در بارهای کاری موثرتر است.

## تست مدل

در این مرحله، عملکرد مدل آموزش‌دیده را بر روی مجموعه‌ای از نقاط داده مجزا به نام داده‌های آزمایشی (که در مرحله دریافت داده تقسیم و نسخه‌سازی شد) ارزیابی می‌شود. استنباط مدل آموزش دیده بر اساس معیارهای انتخاب شده بر اساس مورد استفاده ارزیابی می‌شود. خروجی این مرحله گزارشی از عملکرد مدل آموزش دیده است. در ماژول Deploy، مدل‌های آموزش دیده را به ترتیب در محیط‌های توسعه‌دهنده، تست و تولید مستقر می‌شود. ابتدا با تست اپلیکیشن (که در محیط‌های توسعه دهنده و تست انجام می‌شود) شروع می‌شود.

## تست برنامه

قبل از استقرار یک مدل یادگیری ماشینی جهت تولید، تست استحکام، مقیاس پذیری و امنیت مدل بسیار مهم است. از این رو، ما مرحله "تست برنامه" را داریم که در آن تمام مدل‌های آموزش دیده و برنامه کاربردی را در یک محیط تولید مانند به نام محیط تست یا مرحله بندی به طور دقیق تست می‌شود. در این مرحله، ما ممکن است تست‌هایی مانند تست‌های A/B، تست‌های یکپارچه سازی، تست‌های پذیرش کاربر (UAT) ، یا تست بار را انجام دهیم.

# آموزش و تحویل

* برگزاری جلسات آموزش مفاهیم کلی و نصب و مدیریت بستر Mlops
* تحویل تمامی مستندات مربوط به فاز توسعه
* تاییدیه نماینده فنی سحاب در خصوص تایید صحت مستندات تولید شده
* تاییدیه نماینده فنی سحاب در خصوص تکمیل فاز آموزش
* روش تحویل مستندات
* روش تحویل بستر Mlops

# نصب و راه اندازی MLOPs

* ارائه مستنداتی پیرامون مراحل نصب و راه اندازی کلیه مولفه‌های اصلی بستر Mlops
* به روزرسانی مستندات به ازای هر نسخه از مولفه‌های اصلی بستر Mlops
* کلیه مراحل مندرج در مستندات توسط تیم سحاب با حضور نماینده شرکت آدین، اجرا و صحت سنجی خواهد شد

# منابع موردنیاز

# زمان‌بندی و شکست کارها

پیاده‌سازی MLOps معمولاً شامل مراحل زیر است:

* توسعه مدل: با استفاده از ابزارهایی مانند Jupyter Notebook، TensorFlow، PyTorch و غیره، مدل‌های یادگیری ماشین را توسعه و آموزش دهید.
* کنترل نسخه مدل: نسخه‌های مدل را با استفاده از سیستم‌های کنترل نسخه مانند Git ذخیره و مدیریت کنید.
* تست مدل: مدل‌ها را با استفاده از تست‌های خودکار آزمایش کنید تا مطمئن شوید که الزامات عملکرد و دقت را برآورده می‌کنند.
* استقرار مدل: مدل‌ها را با استفاده از زیرساخت‌ها به عنوان ابزارهای کد مانند Terraform یا CloudFormation در تولید مستقر کنید.
* مانیتورینگ مدل: مدل‌های در حال تولید را با استفاده از ابزارهایی مانند TensorFlow Serving، Seldon یا KFServing جهت شناسایی و پاسخگویی به مسائل نظارت کنید.
* تعمیر و نگهداری مدل: به طور مرتب مدل‌ها را به روز رسانی و نگهداری کنید تا عملکرد آنها را بهبود بخشید و مشکلاتی را که پیش می‌آید برطرف کنید.
* همکاری: تشویق به همکاری بین دانشمندان داده، توسعه دهندگان و تیم‌های عملیاتی جهت اطمینان از MLOهای روان و مؤثر.
* یکپارچه سازی مداوم و استقرار مستمر (CI/CD): فرآیند یکپارچه سازی، آزمایش و استقرار مدل‌ها را با استفاده از ابزارهایی مانند جنکینز یا GitLab به صورت خودکار انجام دهید.
* قابلیت توضیح مدل: اطمینان حاصل شود که مدل‌ها شفاف و قابل توضیح هستند، به طوری که تصمیمات آنها قابل درک و اعتماد توسط ذینفعان باشد.

مراحل اجرای Kubeflow معمولاً شامل موارد زیر است:

* آماده سازی: جمع‌آوری نیازمندی‌ها، ارزیابی زیرساخت‌ها و تعیین محدوده پروژه اجرای Kubeflow.
* نصب: Kubeflow را روی زیرساخت هدف، از جمله خوشه‌های Kubernetes و هر گونه وابستگی لازم، نصب و پیکربندی کنید.
* آماده‌سازی داده‌ها: داده‌ها را جهت استفاده در Kubeflow آماده و سازماندهی کنید، از جمله داده‌ها، پاک‌سازی و عادی‌سازی.
* توسعه مدل: مدل‌های یادگیری ماشین را با استفاده از ابزارهای داخلی Kubeflow یا ادغام با سایر چارچوب‌های ML مانند TensorFlow، PyTorch و Jupyter توسعه دهید.
* آموزش مدل: مدل‌های قطار را روی داده‌های مقیاس بزرگ با استفاده از قابلیت‌های کوچک‌سازی Kubeflow، از جمله آموزش توزیع‌شده و شتاب GPU، آموزش دهید.
* استقرار مدل: با استفاده از ابزارهای استقرار Kubeflow، از جمله TensorFlow Serving، Seldon، و KFServing، مدل‌ها را جهت تولید مستقر کنید.
* نظارت و نگهداری: مدل‌های در حال تولید، از جمله عملکرد، دقت و جابجایی داده‌ها را پایش کنید و مدل‌ها را در صورت لزوم به‌روزرسانی کنید.
* همکاری: تشویق به همکاری بین دانشمندان داده، توسعه دهندگان و تیم‌های عملیاتی جهت اطمینان از MLOهای روان و مؤثر.
* یکپارچه سازی مداوم و استقرار مستمر (CI/CD): فرآیند یکپارچه سازی، آزمایش و استقرار مدل‌ها را با استفاده از ابزارهایی مانند جنکینز یا GitLab به صورت خودکار انجام دهید.
* بررسی و بهینه‌سازی: به طور مداوم پیاده‌سازی Kubeflow از جمله عملکرد، مقیاس‌پذیری و ادغام با ابزارها و خدمات دیگر را بررسی و بهینه کنید.

# تحلیل هزینه

1. Model drift [↑](#footnote-ref-1)
2. <https://www.samsungsds.com/vn/ai-kubeflow/kubeflow.html> [↑](#footnote-ref-2)