

پایتون مقیاس پذیر برای مهندسی داده

Scaling Python for Data Engineering

آموزش پردازش داده‌های بزرگ‌تر از حافظه با پایتون؛ شامل مدیریت مصرف حافظه، بارگذاری هدفمند، پردازش تکه‌ای و موازی سازی، همراه با تمرین روی داده‌های واقعی تاکسی نیویورک.

سطح دوره	مدت پیشنهادی	بخش و درس	شیوه‌ی برگزاری
مقدماتی تا متوسط	۲ روز (۱۶ ساعت)	۸ بخش، ۳۶ درس	حضوری یا آنلاین

پیش نیازها

- تسلط بر پایتون
- آشنایی کاربردی با مبانی pandas: خواندن CSV، انتخاب ستون، گروه بندی و تجمیع

ابزارها و فناوری‌ها

NumPy pandas Python 3
PyArrow

پروژه‌ی عملی

یک جریان پردازش برای داده‌های بزرگ‌تر از حافظه روی داده‌های تاکسی نیویورک

— آنچه شرکت کنندگان می‌آموزند

- ◆ اندازه گیری هزینه‌ی واقعی داده در حافظه و کاهش آن
- ◆ با انتخاب نوع داده‌ی مناسب
- ◆ بارگذاری فقط داده‌ی مورد نیاز و پردازش تکه تکه‌ی داده‌های بزرگ‌تر از حافظه
- ◆ استفاده از NumPy و pandas در مقیاسی که کد ساده دیگر جواب نمی‌دهد
- ◆ موازی سازی پردازش روی چند هسته‌ی پردازنده
- ◆ انتخاب ساختمان‌های داده‌ای که جریان‌های واقعی بر پایه‌ی آن‌ها ساخته می‌شوند

— سرفصل‌های دوره

۱ وقتی داده از حافظه بزرگ‌تر می‌شود	۵ درس
۱. دیوار حافظه	۴. بارگذاری فقط داده‌ی مورد نیاز
۲. سنجش مصرف حافظه و انواع داده	۵. پروژه‌ی راهنما: بررسی داده‌های تاکسی نیویورک
۳. جعبه ابزار مهندس داده	

۱. چرا NumPy
۲. نوع داده، دقت و حافظه
۳. برداری‌سازی: از حلقه تا عملیات آرایه‌ای
۴. انتشار (Broadcasting)
۵. پروژه‌ی راهنمادار: شاخص‌های سفر با NumPy

۱. کاهش اندازه‌ی نوع داده‌های عددی
۲. داده‌های دسته‌ای (Categorical)
۳. عملیات کارآمد و پرهیز از کپی‌های اضافه
۴. گروه‌بندی و تجمیع سریع
۵. پروژه‌ی راهنمادار: بارگذار کم‌مصرف داده‌های تاکسی

۱. ایده‌ی تکه‌بندی داده
۲. تجمیع جریانی
۳. SQLite به‌عنوان انبار داده‌ی محلی
۴. جریان‌های پردازش افزایشی و قابل ازسرگیری
۵. پروژه‌ی راهنمادار: تجمیع داده‌ی چندماهه‌ای که در حافظه جا نمی‌شود

۱. چرا موازی‌سازی (و قفل سراسری مفسر)
۲. multiprocessing و Process Pool
۳. الگوی MapReduce
۴. پردازش موازی تکه‌های داده
۵. پروژه‌ی راهنمادار: تجمیع موازی همراه با اندازه‌گیری

۱. دیکشنری و مجموعه: جست‌وجوی $O(1)$ و حذف تکراری‌ها
۲. پشته و صف
۳. هیپ و صف اولویت
۴. انتخاب ساختار مناسب
۵. پروژه‌ی راهنمادار: ایندکس درون‌حافظه‌ای و Top-N

۱. بازگشت برای پردازش داده
۲. درخت جست‌وجوی دودویی
۳. درخت برای مرتب‌سازی و ادغام
۴. ایندکس‌گذاری یک فایل
۵. پروژه‌ی راهنمادار: ایندکس بازه‌ی زمانی روی داده‌های تاکسی

۱. پروژه پایانی: جریان تحلیلی برای داده‌های
بزرگ‌تر از حافظه

درخواست برگزاری اختصاصی این دوره برای سازمان شما

ثبت درخواست این دوره

lotfinejad.ir/contact

سرفصل، مدت، سطح و تمرین‌های دوره بر اساس نیاز تیم شما تنظیم می‌شود و آموزش به زبان فارسی، به صورت حضوری یا آنلاین برگزار می‌شود. برای ثبت درخواست، فرم مربوط به همین دوره را تکمیل کنید.