مدرس: حسن احمدخانی

نام دوره: دوره آنلاین Data Lakehouse مقدماتی

طول دوره: ۵۰ ساعت

نوع ارائه: آنلاین

نوع ارائه: ارائه در پلیر اختصاصی

پیش نیاز شرکت در دوره: دانش SQL و تجربه کار با یک پایگاه داده RDBMS

مدرس دوره

حسن احمدخانی

» معمار کلان داده در ایرانسل از ۱۴۰۰ تا کنون.

» مشاور مهندسی داده در تکنولایف از ۱۴۰۳.

» مشاور مهندسی داده در زپ از ۱۴۰۲.

» مدرس کلان داده و NoSQLها از سال ۱۳۹۷ تاکنون.

» معمار کلان داده در داتین از ۱۳۹۹ تا ۱۴۰۰.

» مهندس ارشد داده و Oracle DBA در شرکت پژواک از ۱۳۹۴ تا ۱۳۹۸.

» مشاور اوراکل در بیمه ایران از ۱۳۹۵ تا ۱۳۹۶.

» برنامه نویس جاوا و مهندس داده در ایزایران از ۱۳۹۳ تا ۱۳۹۴.

» برنامه نویس جاوا در شرکت پژواک از ۱۳۸۹ تا ۱۳۹۳.

مدرس: حسن احمدخانی

فصل اول : ورود به دنیای Data Lakehouse

این فصل به عنوان دروازه ورود به دنیای مهندسی داده، اهمیت فزاینده داده‌های حجیم و متنوع (بیگ دیتا) و چالش‌های ناشی از ذخیره‌سازی و پردازش آن‌ها را بررسی می‌کند. در این بخش، شما با مسیر شغلی مهندس داده، تفاوت‌های آن با نقش‌هایی مانند DBA و DevOps، و نقش حیاتی این تخصص در عصر هوش مصنوعی آشنا خواهید شد. همچنین، مفاهیم کلیدی مانند ذخیره‌سازی و پردازش توزیع‌شده و مقایسه مدل‌های ذخیره‌سازی سطری و ستونی و مدل‌های پردازشی Batch و Streaming به منظور ایجاد یک دیدگاه جامع از اکوسیستم داده‌های مدرن، معرفی می‌شوند.

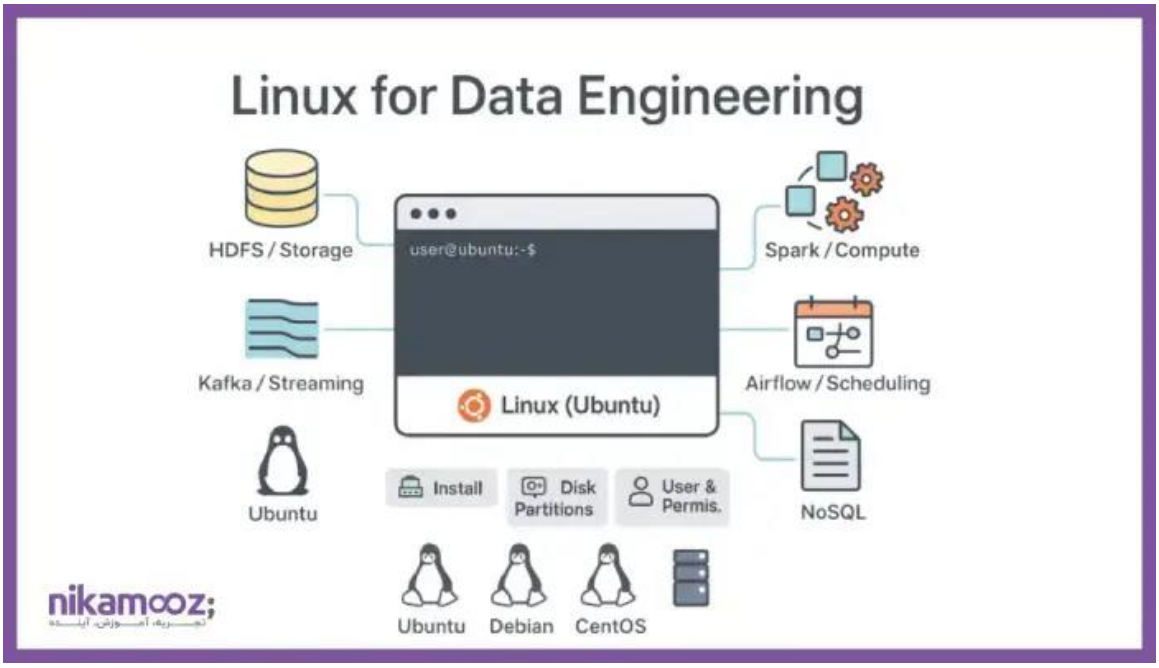
- مروری بر اهمیت داده، پایگاه داده و سیستم های ذخیره سازی
- مهمترین نیازمندی های لازم برای متخصصان مهندسی داده: مسیر حرفه ای شدن
- یک روز کاری مهندس داده چگونه است
- تفاوت ها و شباهت های نقش های مهندسی داده و DBA و DevOps
- عصر AI و اهمیت مهندسی داده
- بیگ دیتا چیست و چه کسانی بیگ دیتا دارند؟
- چالش های داده های حجیم و متنوع و بیگ دیتا
- ذخیره سازی توزیع شده، اهمیت و چالش های آن
- پردازش توزیع شده، اهمیت و چالش های آن
- مدل های ذخیرهازی سطری و ستونی و مثال هایی از آن

مدرس: حسن احمدخانی

فصل دوم: لینوکس برای مهندسان داده

از آنجایی که اغلب پلتفرم‌ها و ابزارهای بیگ دیتا و پردازش توزیع‌شده بر روی سیستم‌عامل لینوکس اجرا می‌شوند، تسلط بر آن برای مهندسان داده ضروری است. این فصل اهمیت لینوکس را در زیرساخت‌های داده‌ای توضیح می‌دهد و بر روی مهم‌ترین توزیع‌ها (مانند اوبونتو) متمرکز می‌شود. شما مهارت‌های عملی مورد نیاز برای کار با لینوکس، از جمله نصب، مدیریت اولیه توزیع‌های مختلف، تنظیمات شبکه و درک مفاهیم پارتیشن‌بندی و مدیریت دیسک را برای آماده‌سازی محیط‌های کاری داده‌محور، فرا خواهید گرفت.

- اهمیت لینوکس
- توزیع های مهم لینوکس
- نصب و کار با Ubuntu
- مروری بر تنظیمات شبکه، پارتیشن بندی و دیسک



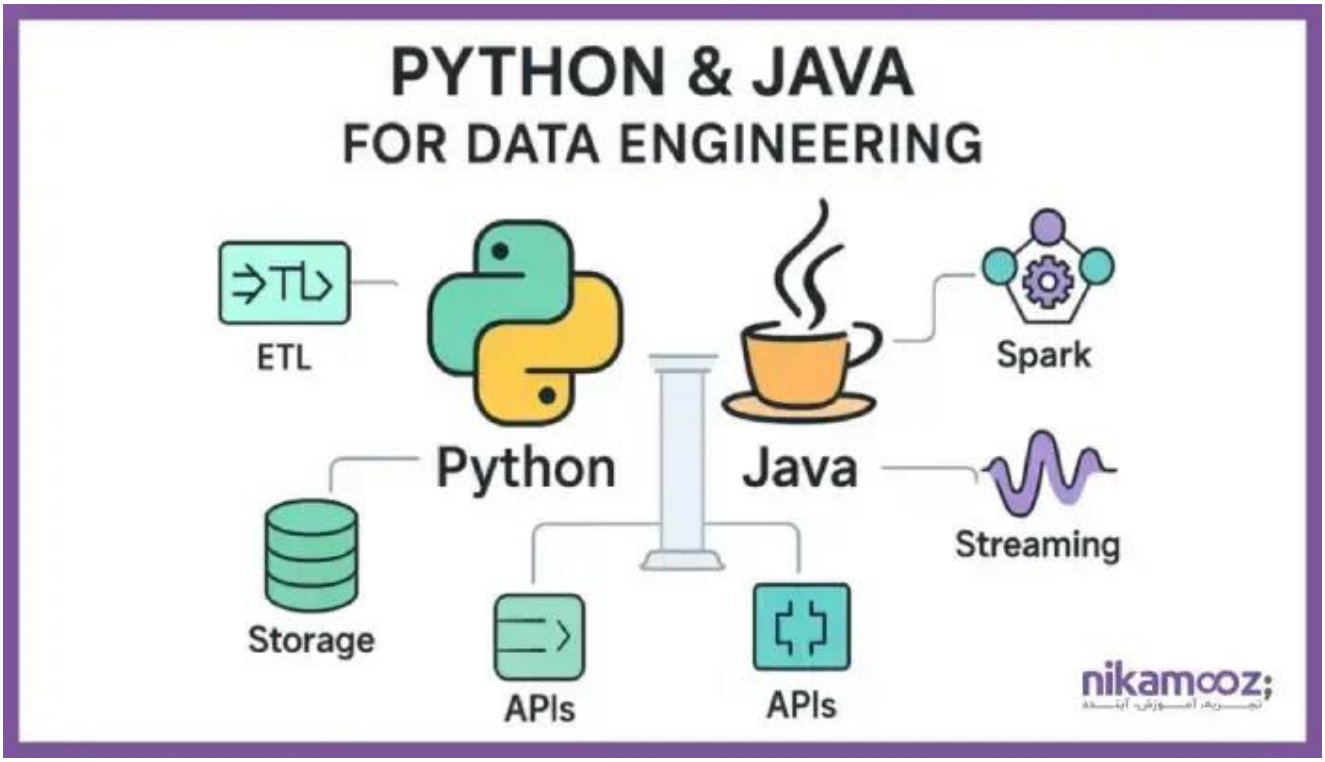
فصل سوم: مقدمه ای بر جاوا و پایتون برای مهندسان داده

این سرفصل به معرفی و بررسی دو زبان برنامه‌نویسی پایتون و جاوا به عنوان ستون‌های اصلی ابزارهای مهندسی داده می‌پردازد. این بخش بر تفاوت مهارت‌های برنامه‌نویسی مورد نیاز مهندس داده در مقایسه با مهندس نرم‌افزار تأکید دارد، در حالی که نکات و اصطلاحات مهم طراحی نرم‌افزار را جهت توسعه کدهای پایدار و بهینه، آموزش می‌دهد. بخش عملی شامل کدنویسی و اجرای سناریوهای رایج پردازش داده با هر دو زبان و همچنین استفاده از گیت برای کنترل نسخه کدها و مدیریت پروژه‌ها خواهد بود تا شرکت‌کنندگان آمادگی لازم برای کار با فریم‌ورک‌های مختلف داده‌ای را کسب کنند.

- زبان های برنامه نویسی مهم برای مهندسان داده
- آیا مهندس داده باید به خوبی یک مهندس نرم افزار در طراحی و برنامه نویسی باشد؟

مدرس: حسن احمدخانی

- نکات و اصطلاحات مهم طراحی و توسعه نرم افزار برای مهندسان داده
- کار با گیت
- کد نویسی و انجام سناریوهای مختلف پردازش داده با پایتون
- کد نویسی و انجام سناریوهای مختلف پردازش داده با جاوا



فصل چهارم: پایگاه داده ها و پلتفرم های NoSQL و Stream Storage

این فصل بر تنوع و اهمیت پایگاه‌های داده NoSQL تمرکز دارد، ویژگی‌های آن‌ها را در مقابل مدل‌های سنتی ACID بررسی کرده و مفاهیمی مانند BASE را توضیح می‌دهد. بخش عملی شامل کار با پایگاه داده MongoDB و اصول مدل‌سازی داده در آن است. در ادامه، مبحث داده‌های جریانی (Stream Data) و چالش‌های آن معرفی می‌شود و پلتفرم Apache Kafka به عنوان ابزار اصلی ذخیره‌سازی جریانی، ساختار، کاربردها و مؤلفه‌های آن (مانند Kafka Connect) مورد بررسی قرار می‌گیرد. در نهایت، معماری و نحوه کار با Elasticsearch برای مدیریت و جستجوی داده‌های حجیم به عنوان یک ابزار قدرتمند تحلیلی، آموزش داده می‌شود.

- اهمیت، ویژگی ها و علت تنوع پایگاه داده های NoSQL
- ویژگی های ACID و BASE در پایگاه داده های NoSQL
- کار با MongoDB
- مدل سازی داده در مانگو دی بی و نکات مدل سازی
- داده های جریانی، ویژگی ها و چالش های آن
- معرفی Apache Kafka و کاربرد و ویژگی های آن
- توزیع های مختلف کافکا و پلتفرم های سازگار با Protocol کافکا

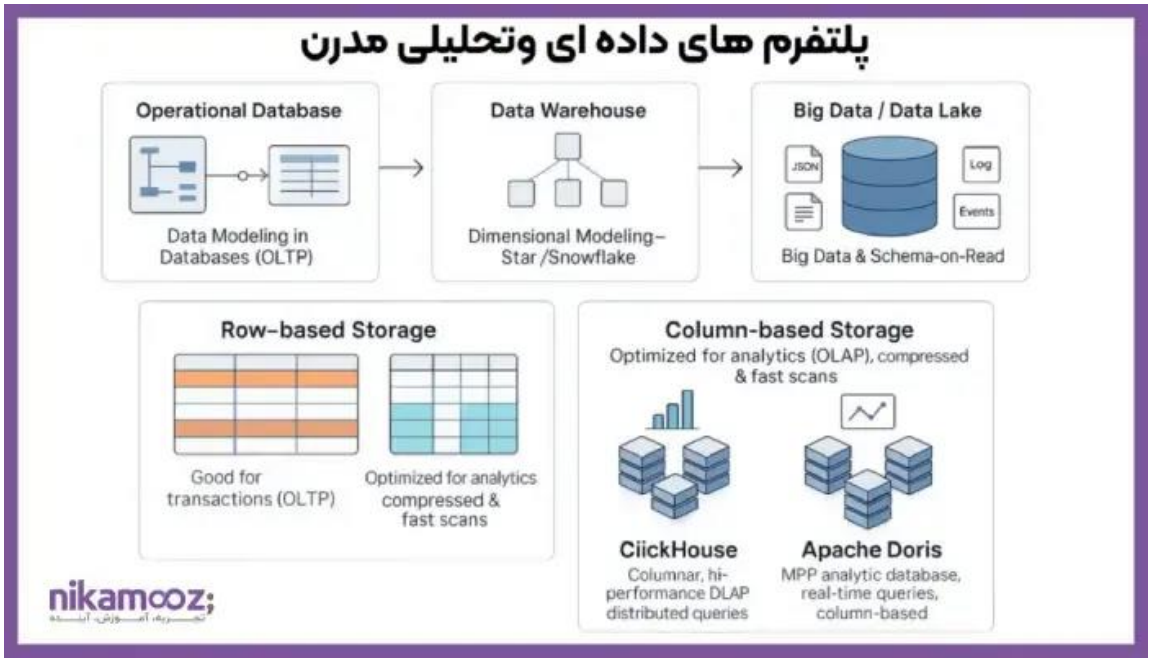
مدرس: حسن احمدخانی

- کار با کافکا و کامپوننت های کافکا مانند Kafka Connect
- اکوسیستم کافکا و انجام سناریو های مختلف ذخیره و بازیابی داده در کافکا
- فرمت های مختلف فایل برای کافکا
- Queue در کافکا
- کاربرد Elasticsearch و معماری آن
- کار با Elasticsearch
- کلاستر Elasticsearch و ساختار آن

فصل پنجم: پلتفرم های داده ای و تحلیلی مدرن

این بخش به معرفی و کاربرد پلتفرم‌های تحلیلی و انبارهای داده مدرن که برای پردازش و تحلیل کارآمد داده‌های حجیم طراحی شده‌اند، می‌پردازد. ابتدا مفاهیم مدل‌سازی داده در انبار داده و تفاوت‌های کلیدی بین ذخیره‌سازی Row-base و Column-base که سرعت کوئری‌ها را تحت تأثیر قرار می‌دهند، مرور می‌شوند. سپس، شرکت‌کنندگان با پلتفرم‌های ستونی با کارایی بالا مانند ClickHouse و Apache Doris آشنا می‌شوند و نحوه کار، ویژگی‌ها و مزایای استفاده از آن‌ها در محیط‌های تحلیلی پیچیده برای ارائه گزارش‌گیری‌های سریع و پاسخ‌گویی به کوئری‌های تحلیلی را فرا می‌گیرند.

- مدل سازی داده در دیتابیس، انبار داده و بیگ دیتا
- تفاوت ذخیره سازی Row-base و Column-base
- کار با ClickHouse و ویژگی های کلیک هاس
- کار با Apache Doris و ویژگی های دوریس



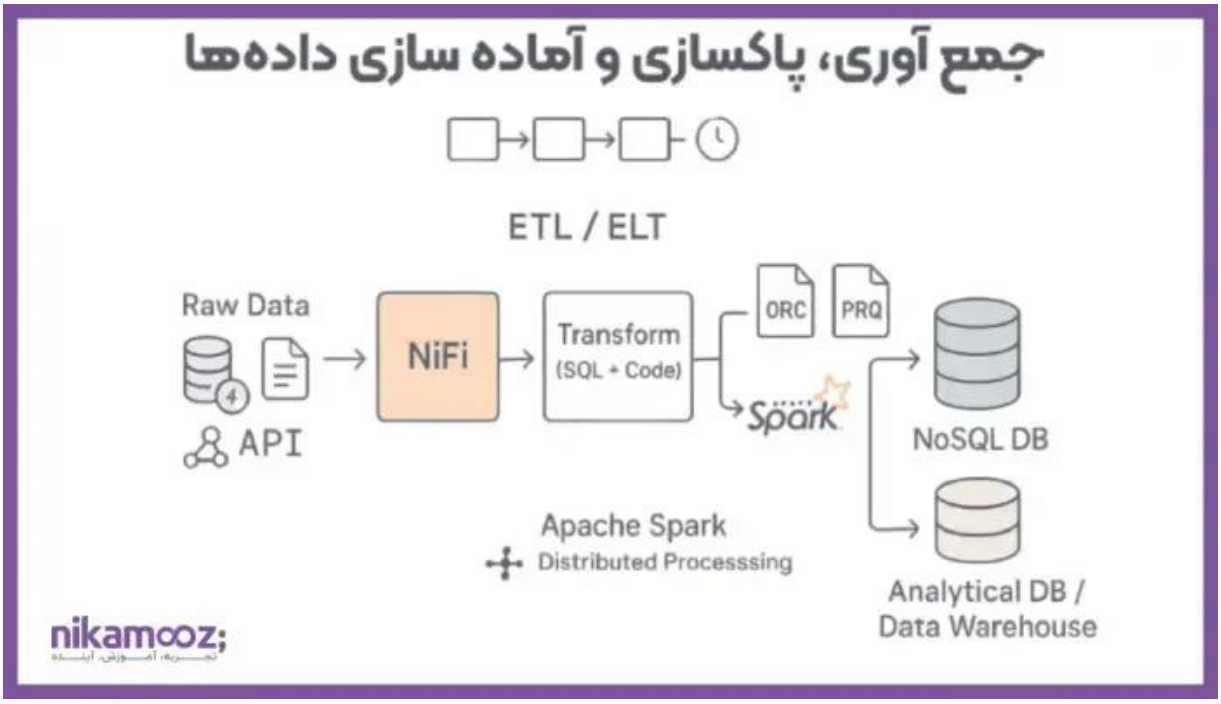
فصل ششم: جمع آوری، پاکسازی و آماده سازی داده ها

این فصل هسته اصلی کار یک مهندس داده را تشکیل می‌دهد و فرآیندهای حیاتی ETL و ELT و اهمیت کیفیت داده را پوشش می‌دهد. این بخش بر روی پیاده‌سازی عملی سناریوهای آماده‌سازی داده‌ها با استفاده از دستورات SQL و یک زبان برنامه‌نویسی تمرکز می‌کند. همچنین، استفاده از ابزارهای خاصی مانند Apache NiFi برای طراحی Data Flowهای جمع‌آوری داده و Apache Spark به عنوان یک فریم‌ورک پردازش توزیع‌شده مقیاس‌پذیر برای اجرای سناریوهای مختلف Spark SQL و خواندن/نوشتن در منابع داده‌ای NoSQL و تحلیلی آموزش داده می‌شود. در پایان، مفاهیم مربوط به فرمت‌های بهینه ذخیره‌سازی مانند ORC و Parquet و روش‌های Scheduling تسک‌های داده‌ای با استفاده از ابزارهایی مانند Airflow بررسی خواهند شد.

- ETL و ELT
- کیفیت داده و معیار های آن
- پیاده سازی سناریو های مختلف آماده سازی داده به کمک یک زبان برنامه نویسی و دستورات SQL
- مقیاس پذیری درلایه پردازش و استفاده از روش های پردازش توزیع شده
- استفاده از Apache NiFi برای طراحی Data Flow های جمع آوری داده
- انجام سناریو های متنوع Ingestion در NiFi به پایگاه داده های NOSQL و تحلیلی
- مروری بر Apache Spark و قابلیت های آن
- انجام سناریو های مختلف پردازش داده در Spark SQL
- معرفی Data Lake و ویژگی‌های آن
- بررسی فرمت فایل های ORC و Parquet و مزایا و معایب آنها
- خواندن و نوشتن در منابع داده ای NoSQL و تحلیلی با آپاچی اسپارک
- روش های Scheduling تسک ها

مدرس: حسن احمدخانی

- استفاده از Airflow برای Scheduling



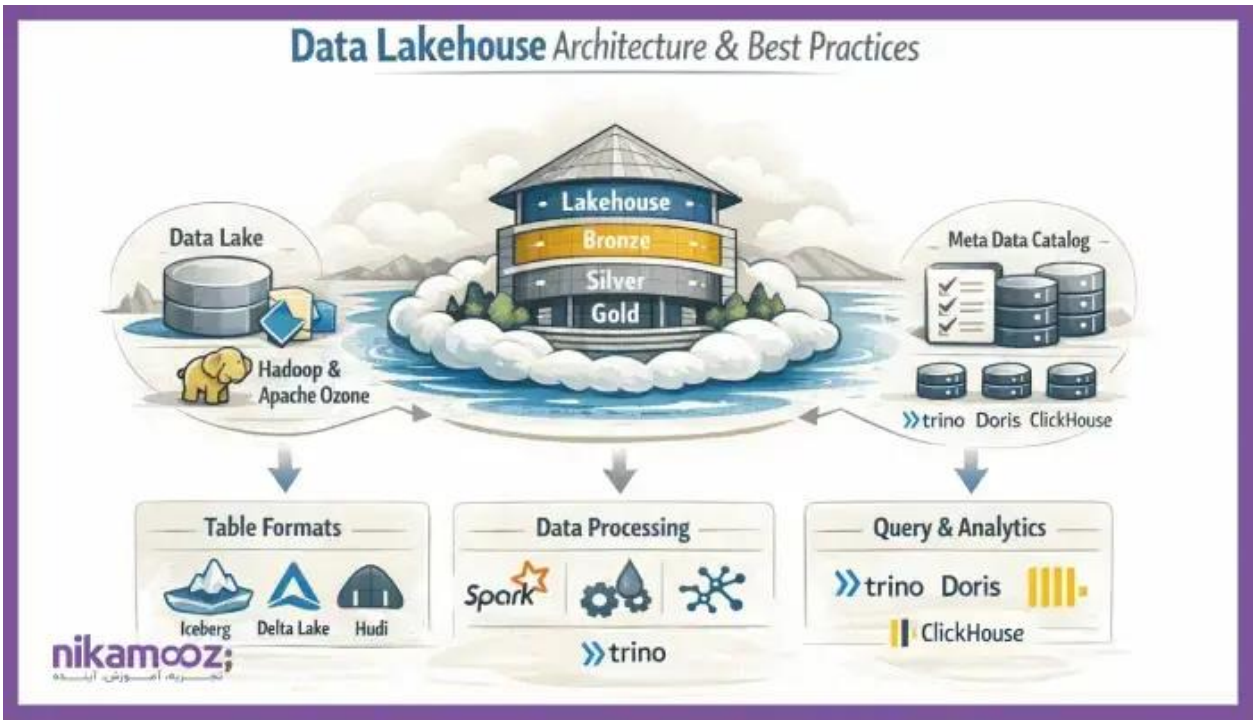
فصل هفتم: ذخیره‌سازی و پردازش کلان داده با روش‌های مدرن در Data Lakehouse

این فصل به‌صورت متمرکز به معماری مدرن Data Lakehouse و نحوه پیاده‌سازی عملی آن می‌پردازد و فرآیند ذخیره‌سازی و پردازش کلان‌داده را در یک بستر یکپارچه آموزش می‌دهد. در این بخش، راه‌اندازی یک Lakehouseبر اساسBest Practice ها، نوشتن داده با استفاده از ابزارهایی مانندSpark ، Trinoو NiFi و همچنین کوئری‌گرفتن و تحلیل داده‌ها با موتورهایی مانندTrino ، Dorisو ClickHouse به‌صورت عملی بررسی می‌شود. همچنین نقشTable Format ها، Data CatalogوQuery Engine ها در ایجاد یک زیرساخت تحلیلی مقیاس‌پذیر و قابل‌اعتماد تبیین خواهد شد.

- نگاهی عمیق به Data Lake و محدودیت‌های آن
- معرفی معماری Medallion و بررسی لایه های آن
- پیاده سازی سناریو های مختلف داده ای روی Data Lake در معماریMedallion
- معرفی Data Lakehouse و مزایای آن
- معرفی و ماهیتTable Format ها و مقایسه باFile Format ها
- بررسی عمیق Iceberg و ویژگی های آن
- پیاده‌سازی سناریوهای مختلف داده‌ای رویData Lakehouse
- پیاده سازی سناریو های مختلف داده ای روی Data Lakehouse در معماریMedallion

مدرس: حسن احمدخانی

- معرفی Apache Ozone و Hadoop HDFS و کاربردها و ویژگی‌های آن‌ها
- کاربرد Data Catalog و Meta Data Store در معماری Data Lake وLakehouse
- معرفی و استفاده از Trino به عنوان Query Engine برایLakehouse
- استفاده از Doris و ClickHouse به عنوان Query Engine برایLakehouse
- راه‌اندازی یک Data Lakehouse بر اساس Best Practice ها
- نوشتن داده روی Lakehouse به کمک Spark و Trino وNiFi
- کوئری داده‌های Lakehouse به کمک Trino و Doris وClickHouse



سوالات متداول

Data Lakehouse چیست؟

Data Lakehouseیک معماری مدرن داده است که قابلیت‌های Data Lake و Data Warehouse را در یک بستر واحد ترکیب می‌کند. در این مدل، داده‌های حجیم و متنوع به‌صورت مقرون‌به‌صرفه ذخیره می‌شوند و

همزمان امکان انجام تحلیل‌های ساختاریافته، گزارش‌گیری و پردازش‌های پیشرفته روی همان داده‌ها فراهم است. این معماری (با حفظ کیفیت داده، امنیت و قابلیت اطمینان)، زیرساختی یکپارچه و مقیاس‌پذیر برای تحلیل داده و هوش مصنوعی ایجاد می‌کند.

مهندسی داده چیست و چه کسی مهندس داده است؟

منابع موجود به طور مستقیم تعریف جامعی از اینکه مهندسی داده چیست یا دقیقاً چه کسی مهندس داده است، ارائه نمی‌دهند. با این حال، این حوزه شامل بررسی اهمیت فزاینده داده‌های حجیم و متنوع (بیگ دیتا)، چالش‌های ذخیره‌سازی و پردازش آن‌ها و نقش حیاتی این تخصص در عصر هوش مصنوعی است.

آینده شغلی مهندسی داده به کدام سمت می‌رود؟

با انفجار حجم داده‌ها و ظهور عصر هوش مصنوعی نیاز به مهندسان داده بیش از هر زمان دیگری احساس می‌شود. آینده این شغل به سمت اتوماسیون فرآیندهای داده، استفاده از پلتفرم‌های ابری و پردازش‌های آنی (Real-time)در حرکت است. شرکت‌های بزرگ جهان و ایران به شدت به دنبال متخصصانی هستند که بتوانند چالش‌های بیگ‌دیتا (Big Data) و پردازش‌های توزیع‌شده را مدیریت کنند.

چه پیش‌نیازهایی برای شرکت در این دوره نیاز است؟

برای ورود به این دوره، داشتن دانش کافی در زمینه SQLو تجربه کار با حداقل یکی از پایگاه داده‌های رابطه‌ای (RDBMS)ضروری است. از آنجایی که در طول دوره به مباحث برنامه‌نویسی و محیط‌های عملیاتی پرداخته می‌شود. آشنایی اولیه با منطق برنامه‌نویسی به شما کمک می‌کند تا از فصل‌های مربوط به پایتون، جاوا و لینوکس بیشترین بهره را ببرید.

مهندس داده با مهندس نرم افزار چه تفاوتی دارد و چه پروژه‌هایی این دو اجرا می‌کنند؟

این دوره بر تفاوت مهارت‌های برنامه‌نویسی مورد نیاز مهندس داده در مقایسه با مهندس نرم‌افزار تأکید دارد. همچنین، نکات و اصطلاحات مهم طراحی نرم‌افزار جهت توسعه کدهای پایدار و بهینه آموزش داده می‌شود. مهندسان داده در این دوره، اجرای سناریوهای مختلف آماده‌سازی داده به کمک زبان برنامه‌نویسی و دستورات SQL و انجام سناریوهای متنوع Ingestion با ابزارهایی مانند Apache NiFi را می‌آموزند.

در اکوسیستم مهندسی داده، Apache Spark چیست و چه کاربردی دارد؟

Apache Spark به عنوان یک فریم‌ورک پردازش توزیع‌شده مقیاس‌پذیر معرفی شده است. کاربرد آن شامل اجرای سناریوهای مختلف Spark SQL و همچنین خواندن و نوشتن در منابع داده‌ای NoSQL و تحلیلی است.

در این دوره کدام یک از NoSQLها کاربرد دارد؟

در این دوره، به صورت تخصصی بر روی دو مورد از محبوب‌ترین و پرکاربردترین پایگاه‌های داده غیررابطه‌ای تمرکز می‌کنیم:

مدرس: حسن احمدخانی

- **MongoDB:** برای مدیریت داده‌های سندمحور.(Document-based)
- **Elasticsearch:** به عنوان یک موتور جستجو و تحلیل قدرتمند برای داده‌های حجیم. همچنین تفاوت‌های ساختاری ACID و BASE در این سیستم‌ها به طور کامل بررسی می‌شود.

کافکا چیست؟ آیا در این دوره Kafka تدریس می‌شود؟

بله، Apache Kafka به عنوان پلتفرم اصلی برای ذخیره‌سازی داده‌های جریانی (Stream Storage) معرفی شده است. در این دوره ساختار، کاربردها و مؤلفه‌های آن مانند Kafka Connect مورد بررسی قرار می‌گیرد.

ClickHouse چیست و چه کاربردی دارد؟

ClickHouse یکی از پلتفرم‌های ستونی با کارایی بالا است که برای پردازش و تحلیل کارآمد داده‌های حجیم طراحی شده است. کاربرد آن شامل استفاده در محیط‌های تحلیلی پیچیده برای ارائه گزارش‌گیری‌های سریع و پاسخ‌گویی به کوئری‌های تحلیلی است.

تفاوت مدل‌های ذخیره‌سازی سطری (Row-based) و ستونی (Columnar) در چیست و چرا در این دوره روی دیتابیس‌های ستونی تأکید شده است؟

در مدل ذخیره‌سازی سطری، داده‌ها به‌صورت رکوردهای کامل و پیوسته ذخیره می‌شوند که برای تراکنش‌های عملیاتی (OLTP) مناسب است، زیرا بازیابی یا به‌روزرسانی یک رکورد کامل سریع انجام می‌شود. در مقابل، در مدل ستونی، مقادیر هر ستون جداگانه ذخیره می‌شوند که امکان فشرده‌سازی بهینه و بازیابی سریع ستون‌های خاص را فراهم می‌کند. این ویژگی باعث می‌شود دیتابیس‌های ستونی برای پردازش تحلیلی (OLAP) و کوئری‌های تجمیعی ایده‌آل باشند، چرا که تنها داده‌های مرتبط بارگذاری شده و حجم I/O کاهش می‌یابد. در دوره‌های مدرن، با رشد نیاز به تحلیل داده‌های حجیم، دیتابیس‌های ستونی مانند ClickHouse و Apache Dorisبه‌دلیل کارایی بالا در پردازش کوئری‌های پیچیده و گزارش‌گیری، مورد تأکید قرار گرفته‌اند.

منظور از پردازش توزیع‌شده چیست و چگونه در لایه Ingestion و Processing پیاده‌سازی می‌شود؟

پردازش توزیع‌شده به تقسیم وظایف پردازشی بین چندین گره (Node) در یک شبکه اشاره دارد که امکان مقیاس‌پذیری، تحمل خطا و کارایی بالا را فراهم می‌کند. در لایهIngestion ، این مفهوم با ابزارهایی مانند Apache Kafkaپیاده‌سازی می‌شود که داده‌ها را به‌صورت جریانی و موازی بین چندین پارتیشن توزیع می‌کند. در لایهProcessing ، فریم‌ورک‌هایی مانند Apache Spark از پردازش توزیع‌شده برای اجرای عملیات ETL/ELT استفاده می‌کنند و با تقسیم داده‌ها به بلوک‌های کوچک، پردازش موازی روی کلاسترها را ممکن می‌سازند. این معماری چالش‌های حجم بالای داده را با توزیع بار پردازشی حل می‌کند.

آیا خرید اقساطی امکان‌پذیر است؟

بله، امکان خرید اقساطی با اسنپ پی فراهم شده است. برای اطلاعات بیشتر می‌توانید با مشاورین مجموعه در تماس باشید یا راهنمای **خرید اقساطی دوره آموزشی با اسنپ پی** را مطالعه بفرمایید.

مدرس: حسن احمدخانی

آیا فیلم دوره رکورد می گردد؟

بله، دسکتاپ و صدای مدرس رکورد خواهد شد و در پلیر اختصاصی اسپات پلیر به همراه کلید لایسنس ارائه خواهد شد. شما در سیستم عامل‌های ویندوز، اندروید، آیفون (سیب، اناردون)، مک بوک می‌توانید فیلم را مشاهده کنید.

افرادی که بصورت لایو کلاس را مشاهده می‌کنند، آیا امکان پرسش و پاسخ دارند؟

شما بصورت چت آنلاین می توانید سوالات خود را بپرسید و مدرس هم سوالات شما را پاسخ خواهد داد. البته توجه داشته باشید که این پروسه پاسخگویی هر ۴۰ دقیقه یکبار خواهد بود تا مدرس رشته کلام از دستش خارج نشود. البته که گروه تلگرامی دوره در اختیار شما است و می توانید سوالات خود را آنجا هم مطرح کنید.

آیا امکان برگزاری دوره‌های سازمانی وجود دارد؟

بله؛ در نیک آموز امکان برگزاری دوره‌های سازمانی به‌صورت تخصصی فراهم شده است. به منظور ثبت درخواست، کافی است اطلاعات خود و دوره سازمانی مدنظر را در **فرم درخواست آموزش سازمانی** ثبت کنید تا ما با شما تماس بگیریم.

می‌خواهم مشاوره / تدریس خصوصی بگیرم؟

برای اینکه بتوانید، مشاوره / تدریس خصوصی بگیرید، لطفاً **فرم درخواست مشاوره مدرسین** را تکمیل نمایید تا کارشناسان ما با شما تماس بگیرند.

من باز هم سؤال دارم؛ امکان ارتباط با مشاوران نیک آموز وجود دارد؟

بله؛ شما می‌توانید از مشاوره‌های **نیک آموز** به‌عنوان راهنما در مسیر خود استفاده کنید. برای این منظور، لطفاً شماره خود را در فرم مشاوره صفحه دوره وارد کنید تا مشاوران نیک آموز با شما تماس بگیرند.

شماره تماس فروش: **۰۲۱۹۱۰۷۰۰۱۷** – داخلی ۱

شماره تماس پشتیبانی: **۰۲۱۹۱۰۷۰۰۱۷** – داخلی ۲

آیا می‌توانم مشاوره سازمانی برای پروژه دریافت کنم؟

شما می‌توانید با مراجعه به **فرم درخواست مشاوره تخصصی**، از متخصصان نیک آموز مشاوره دریافت کنید و با به‌کارگیری مهارت‌های تجربی تیم ما، در ارتباط با پروژه‌های تخصصی خود، راهنمایی دریافت کنید.

نحوه پشتیبانی دوره به چه صورت است؟

رضایت شما از دوره آموزشی و کمک به رفع مشکلات احتمالی برای ما اهمیت زیادی دارد. به همین دلیل، یک گروه پشتیبانی در تلگرام ایجاد شده است تا شما بتوانید درصورت نیاز، مسائل خود را در این بستر مطرح کنید. تا حداکثر ۴۸ ساعت کاری پس از ثبت نام در دوره، با شما تماس گرفته می‌شود و فرآیند عضویت شما در گروه تلگرام نهایی خواهد شد. توجه شود که در آینده، سیستم تیکتینگ راه‌اندازی می‌شود و فرآیند پشتیبانی از گروه تلگرامی به آن جا منتقل خواهد شد.

ثبت نام آنلاین : دوره آنلاین Data Lakehouse مقدماتی