دانشگاه خلیج فارس بوشهر

Data Science

علم داده ها

فرصت ایجاد یک کسب و کار جدید

ارائه:
حسین حق‌بین

مرداد ماه ۱۳۹۹

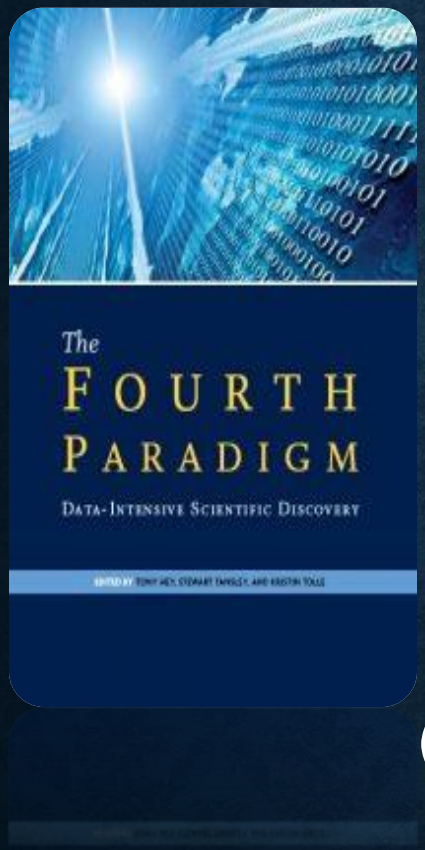
بسم الله الرحمن الرحيم

- ✓ علم داده‌ها چیست؟
- ✓ دانشمند داده به چه کسی گفته می‌شود؟
- ✓ موقعیت‌های شغلی در حوزه علم داده
- ✓ علم داده در استارتاپ‌ها
- ✓ استارتاپ‌های موفق در حوزه علم داده

علم داده‌ها چیست

علم داده‌ها (Data Science)، یک زمینه میان‌رشته‌ای است که از روش‌ها، فرآیندها، الگوریتم‌ها و سیستم‌های علمی برای استخراج دانش (Knowledge) از داده‌های خام (Raw Data) استفاده می‌کند.

در علم داده نظریه‌های علوم گوناگون از جمله آمار، علوم کامپیوتر، ریاضیات، و دیگر روش‌ها مانند یادگیری ماشین، داده‌کاوی و بصری‌سازی داده‌ها مورد استفاده قرار می‌گیرد.



علم داده را پارادایم چهارم علم می‌دانند.

Tansley and Tolle (2009)

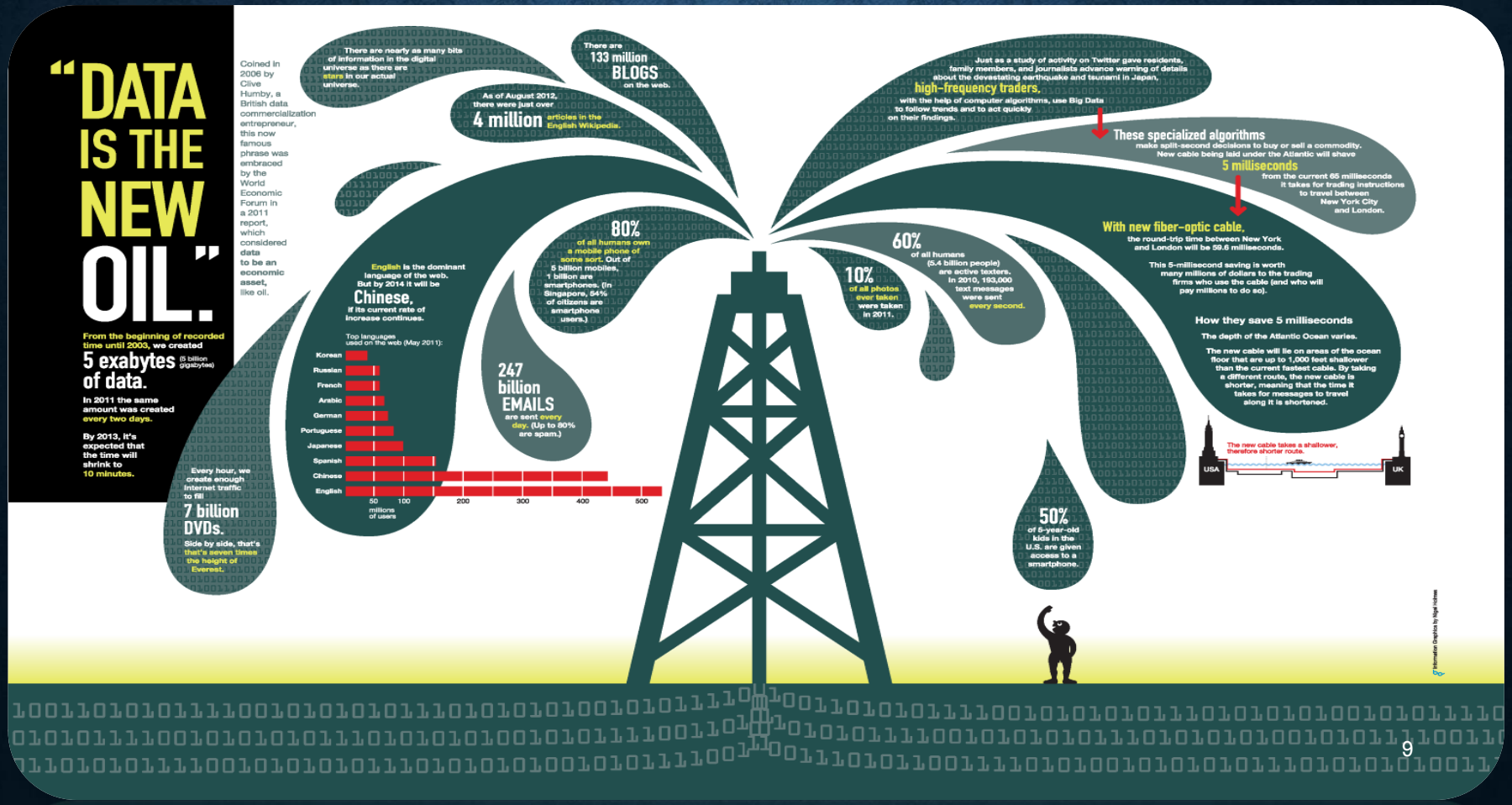
1. شواهد تجربی (Empirical Evidence)
2. نظریه علمی(Scientific Theory)
3. علوم محاسباتی (Computational Science)
4. علم داده (Data Science)

دانشمندان بر این باور هستند که

«کلیه موارد مربوط به علم تحت تاثیر فناوری های داده‌محور در حال تغییر است».

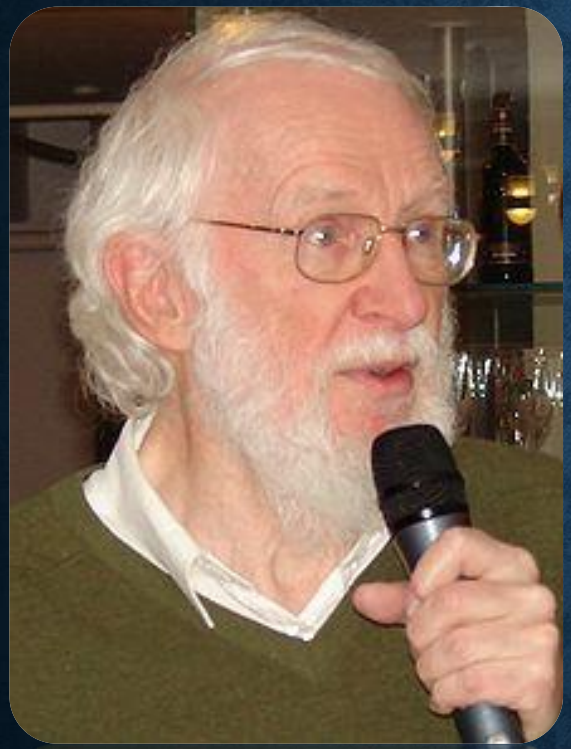
مجمع جهانی اقتصاد در سال ۲۰۱۱ اعلام کرد که:

((داده‌ها نفت عصر جدید هستند))



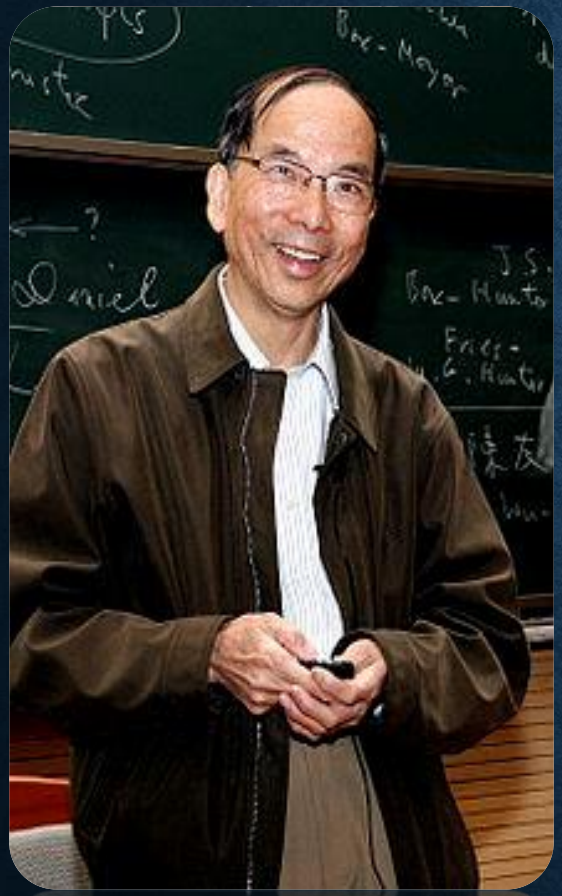
آینده متعلق افراد یا شرکت‌هایی است که بتوانند داده‌ها را به محصول تبدیل کنند.

تاریخچه علم داده



در سال ۱۹۶۰، «پیتر نائور» (Peter Naur) از این عبارت به عنوان جایگزینی برای علوم کامپیوتر استفاده کرد. نائور بعدها اصطلاح «داده‌شناسی» (datalogy) را بدین منظور معرفی کرد. وی در سال ۱۹۷۴ در مقاله‌ای با عنوان «بررسی دقیق روش‌های کامپیوتری» از اصطلاح علم داده برای بیان پردازش‌های داده آن دوران استفاده کرد.

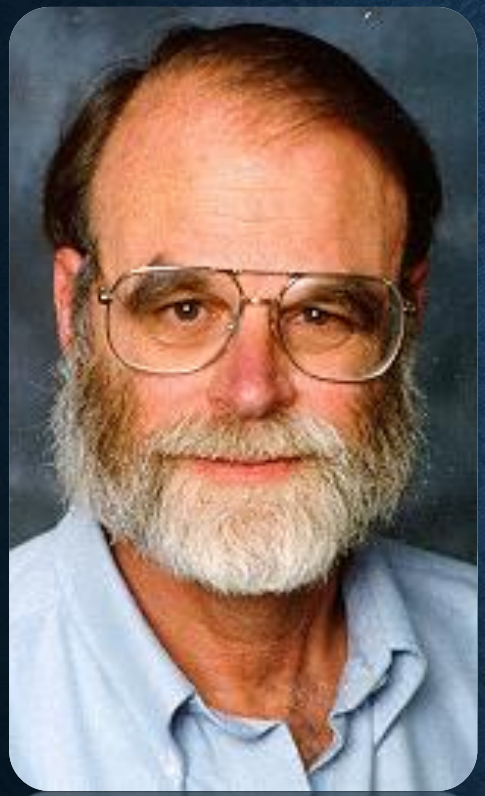
در سال ۱۹۹۶، اعضای «فدراسیون بین‌المللی جامعه دسته‌بندی» (IFCS) برای گردهمایی دو سال یکبار خود، در شهر کوبه ژاپن گردهم آمدند. در گردهمایی مذکور، برای اولین بار از اصطلاح علم داده به عنوان اسم کنفرانس علم داده، دسته‌بندی و روش‌های مرتبط – استفاده شد.



در نوامبر سال ۱۹۹۷، سی اف جف وو (C. F. Jeff Wu) ، آماردان و استاد تحلیلگر شرکت کوکا-کولا، سخنرانی افتتاحیه رویدادی در حوزه علم داده را با عنوان ((آمار = علم داده؟)) در دانشگاه میشیگان انجام داد و از آمار به عنوان علمی که به علم داده و از آماردان به عنوان افرادی که به دانشمندان داده تغییر نام داده‌اند یاد کرد.

در آپریل سال ۲۰۰۲، کمیته داده برای دانش و فناوریِ (CODATA)، انتشار مجله‌ای با عنوان علم داده (Data Science Journal) را آغاز کرد. این اثر، بر مسائلی مانند توصیف سیستم‌های داده، نشر آن‌ها در اینترنت، کاربردها و مسائل قانونی مربوط به این حوزه متمرکز بود.

COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL



در حدود سال ۲۰۰۷، «جیم گری» (Jim Gray)، برنده جایزه تورینگ، علوم داده محور را به عنوان چهارمین پارادایم علم معرفی کرد که از تحلیل محاسباتی داده‌های بزرگ به عنوان روشی علمی جهت ساخت دنیایی که در آن ادبیات علم و همه داده‌های علمی آنلاین هستن استفاده می‌کند.

در سال ۲۰۱۲، دانراجی پاتیل (Dhanurjay Patil ”DJ“)، در مقاله «دانشمند داده: جذاب‌ترین شغل قرن ۲۱» که در مجله بررسی کسب‌وکار هاروارد منتشر شد، از دانشمندان داده به عنوان نژادی جدید یاد کرده که کمبود آن‌ها منجر به محدودیت‌های جدی در برخی از بخش‌های صنعت و دانشگاه می‌شود.

- در سال ۲۰۱۳، اتحادیه اروپایی تحلیل داده (EuADS) در لوکزامبورگ تاسیس شد.
- سال ۲۰۱۵، «ژورنال بین‌المللی علم و تحلیل داده» (International Journal of Data Science and Analytics) توسط اسپرینگر به‌منظور انتشار کارهای انجام پذیرفته در حوزه علم داده و تحلیل داده‌های کلان (مِه داده) بنا شد.

دانشمند داده به چه کسی گفته می‌شود

جف هامربچر (Jeff Hammerbacher) از لیدرهای تیم تحلیل داده فیسبوک می‌نویسد:

یک دانشمند داده در تیم باید بتواند خط لوله پردازش چند مرحله ای در پایتون را بنویسد، یک آزمون فرضیه را طراحی کند ، تجزیه و تحلیل رگرسیون را بر روی نمونه های داده با R انجام دهد ، یک الگوریتم را برای برخی از محصولات یا سرویس‌ها در هادوپ طراحی و پیاده سازی کند یا اینکه نتایج تحلیل‌های ما را به دیگر اعضای سازمان انتقال دهد.

مهارت های اساسی علم داده

1. **تجزیه و تحلیل آماری:** الگوهای موجود در داده ها را مشخص کنید. این شامل داشتن حس شدید تشخیص الگوی و تشخیص ناهنجاری است
2. **یادگیری ماشین:** الگوریتم ها و مدل های آماری را پیاده سازی کنید تا رایانه بتواند به طور خودکار از داده ها یاد بگیرد.
3. **علوم کامپیوتر:** از اصول هوش مصنوعی ، سیستم های پایگاه داده ، تعامل انسان / کامپیوتر ، تجزیه و تحلیل عددی و مهندسی نرم افزار استفاده کنید.

- ✓**برنامه نویسی:** برنامه های رایانه ای را بنویسد و مجموعه داده های بزرگ را برای کشف پاسخ به مشکلات پیچیده آنالیز کند. دانشمندان داده باید بتوانند به راحتی به زبانهای مختلفی مانند **Java**، **R**، **Python** و **SQL** کدنویسی کنند.
- ✓**داستان پردازی داده ها:** با استفاده از داده ها، اغلب برای مخاطبان غیر فنی، بینش‌های عملی را برقرار کند.

موقعیت‌های شغلی در حوزه علم داده

۱. دانشمند داده (DATA SCIENTIST)

- ✓متوسط درآمد سالانه: $139,840
- ✓الزامات شغلی: یافتن، تمیز کردن و سازماندهی داده‌ها برای شرکت ها. دانشمندان داده باید بتوانند مقادیر زیادی از اطلاعات پیچیده و خام را تجزیه و تحلیل کنند تا الگویی را پیدا کنند که به نفع یک سازمان باشد و به تصمیم گیری استراتژیک در زمینه تجارت کمک کند. در مقایسه با تحلیلگران داده، دانشمندان داده بسیار فنی تر هستند.

DATA SCIENTIST

"AS RARE AS UNICORNS"

Languages

R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

Skills & Talents

- ✓ *Distributed computing*
- ✓ *Predictive modeling*
- ✓ *Story-telling and visualizing*
- ✓ *Math, Stats, Machine Learning*

Role

Cleans, massages and organizes (big) data

Mindset

Curious data wizard

HIRED BY

Google Microsoft Adobe

۲. مهندس یادگیری ماشین
(MACHINE LEARNING ENGINEER)

- ✓متوسط درآمد سالانه: ۱۱۴،۸۲۶ دلار
- ✓الزامات شغلی: برنامه‌نویسانی هستند ، که تمرکز آنها بر برنامه‌نویسی هوش مصنوعی برای ماشین است. آنها برنامه‌هایی را ایجاد می کنند که ماشین ها را قادر می سازد به طور هوشمند برای انجام برخی از کارها اقداماتی انجام دهند. نمونه ای از سیستمی که یک مهندس یادگیری ماشین روی آن کار می کند رانندگی یک اتومبیل بدون سرنشین است.آنها به طور معمول به آمار و مهارت های برنامه نویسی قوی و همچنین دانش مهندسی نرم افزار نیاز دارند.

۴. معمار داده
(DATA ARCHITECT)

- متوسط درآمد سالانه: ۱۰۸،۲۷۸ دلار
- الزامات شغلی:. معماران داده چگونگی ذخیره، بازیابی، یکپارچه‌سازی و مدیریت داده‌ها توسط اشخاص و سیستم های IT مختلف را تعریف می‌کنند. همچنین وظیفه مدیریت برنامه‌های کاربردی را که این داده‌ها را به نوعی استفاده یا پردازش می‌کنند، دارند.

DATA ARCHITECT

THE CONTEMPORARY DATA MODELLER

Languages

SQL, XML, Hive, Pig, Spark

Skills & Talents

- ✓ *Data warehousing solutions*
- ✓ *In-depth knowledge of database architecture*
- ✓ *Extraction Transformation and Load (ETL), spreadsheet and BI tools*
- ✓ *Data modeling*
- ✓ *Systems development*

Role:

Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset:

Inquiring ninja with a love for data architecture design patterns

HIRED BY

VISA Coca-Cola logitech

۵. مهندس داده
(DATA ENGINEER)

- متوسط درآمد سالانه: ۱۰۲،۸۶۴ دلار
- الزامات شغلی: پردازش دسته‌ای (batch) یا پردازش زمان-واقعی (real-time) روی داده های جمع آوری شده و ذخیره شده. مهندسان داده همچنین مسئول ساخت و نگهداری خطوط لوله داده (data pipelines) هستند که یک اکوسیستم داده قوی و به هم پیوسته در سازمان ایجاد می کنند و اطلاعات را برای دانشمندان داده در دسترس می کنند.

DATA ENGINEER

SOFTWARE ENGINEERS BY TRADE

Role

Develops, constructs, tests and maintains architectures (such as databases and large-scale processing systems)

Mindset

All-purpose everyman

HIRED BY

Spotify

Languages

SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

Skills & Talents

- ✓ *Database systems (SQL & NO SQL based)*
- ✓ *Data modeling & ETL tools*
- ✓ *Data APIs*
- ✓ *Data warehousing solutions*

۶. توسعه دهنده هوش کسب و کار (BI)
BUSINESS INTELLIGENCE DEVELOPER

- متوسط درآمد سالانه: ۸۱،۵۱۴ دلار
- الزامات شغلی: توسعه دهندگان **BI** استراتژی‌هایی را برای کمک به کاربران مشاغل در یافتن سریع اطلاعات مورد نیاز برای تصمیم گیری در مورد شغل بهتر، طراحی و توسعه می دهند. با درک بسیار بالا از داده‌ها، ابزارهای **BI** استفاده می‌کنند یا برنامه‌های تحلیلی **BI** را توسعه می‌دهند تا درک کاربران نهایی از سیستم خود را تسهیل کنند.

BUSINESS ANALYST

CHANGE AGENT

Languages

SQL

Skills & Talents

- ✓ *Basic tools (e.g. MS Office)*
- ✓ *Data visualization tools (e.g. Tableau)*
- ✓ *Conscious listening and storytelling*
- ✓ *Business Intelligence understanding*
- ✓ *Data modeling*

Role

Improves business processes as intermediary between business and IT

Mindset

Resilient project juggler

HIRED BY

UBER DELL ORACLE

۷. آماردان (STATISTICIAN)

- ✓متوسط درآمد سالانه: ۷۶،۸۸۴ دلار
- ✓الزامات شغلی: آماردان به منظور شناسایی روندها و روابطی که می تواند برای اطلاع رسانی در تصمیم گیری سازمانی مورد استفاده قرار گیرد، جمع آوری، تجزیه و تحلیل و تفسیر داده ها را انجام می دهند. علاوه بر این، مسئولیت‌های روزمره آماردان اغلب شامل فرآیندهای جمع آوری داده های طراحی ، ارتباط یافتن به ذینفعان و مشاوره استراتژی سازمانی است.

STATISTICIAN

'HISTORIC LEADERS OF DATA'

Languages

R, SAS, SPSS, Matlab, Stata, Python, Perl, Hive, Pig, Spark, SQL

Skills & Talents

- ✓ *Statistical theories & methodology*
- ✓ *Data mining & machine learning*
- ✓ *Distributed Computing (Hadoop)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Cloud tools*

Role

Collects, analyzes and interprets- qualitative as well as quantitive data with statistical theories and methods

Mindset

Logical and enthusiastic stats genius

HIRED BY

LinkedIn Johnson-Johnson PEPSICO

۸. تحلیلگر داده (DATA ANALYST)

- ✓**متوسط درآمد سالانه:** ۶۲۴۵۳ دلار
- ✓**الزامات شغلی:** تحلیلگر داده بایستی مجموعه داده های بزرگ را متناسب با تجزیه و تحلیل مورد نظر برای شرکت‌ها تبدیل و دستکاری کند. برای بسیاری از شرکت‌ها ، این نقش همچنین می‌تواند شامل ردیابی تجزیه و تحلیل باشد. تحلیلگران داده همچنین با تهیه گزارشهایی برای رهبران سازمانی که به طور مؤثر روندها و بینش‌های حاصل از تحلیل آنها را بطور مؤثر ارتباط می دهند، به فرآیند تصمیم گیری کمک می کنند.

DATA ANALYST

'DATA DETECTIVE'

Languages

R, Python, HTML, Javascript, C/C++, SQL

Role

Collects, processes and performs statistical data analyses

Skills & Talents

- ✓ *Spreadsheet tools (e.g. Excel)*
- ✓ *Database systems (SQL and NO SQL based)*
- ✓ *Communication & visualization*
- ✓ *Math, Stats, Machine Learning*

Mindset

Intuitive data junkie with high "figure-it-out" quotient

HIRED BY

IBM hp DHL

برنامه جامع برای تبدیل شدن به دانشمند داده

- The most comprehensive Data Science learning plan for 2017

علم داده در استارتاپ‌ها

اولین قدم برای ساختن هر محصول داده‌ای، جمع آوری داده است.

برای جمع آوری داده‌ها ، ابتدا باید تصوری از کاربران پایه و تعداد لاگ رویدادهایی که به برنامه شما دسترسی دارند، باشد.

مثال: فرض کنید قصد دارید یک محصول نرم‌افزاری را بفروشید. شما نیاز دارید اطلاعاتی پایه درباره کاربران محصول خود در مراحل مختلف نصب، استفاده و خرید داشته باشید.

برای انجام این کار می‌توان برنامه را با ردیاب‌ها (Trackers) جاسازی کرد. قرار دادن رویدادهای ردیابی (Tracking Events) در سمت-کاربر (Client-side) ضروری است زیرا داده‌هایی را به سرور ارسال می کند که برای توسعه محصولات کاربرد دارد.

انواع داده‌های یک استارتاپ

Raw Data

- Do not have any schema
- Not present in any designated format
- Example: tracking events

Processed Data

- Implementation of schemas over the raw data
- Encoded in specified formats

Cooked Data

- Summary of the processed data

نکاتی درباره راه اندازی یک استارتاپ در حوزه علوم داده

- ✓کاربران را در هر فاز پروژه همراه خود کنید.
- ✓برای جلوگیری از کار مضاعف، مهندسان داده را با دانشمندان داده- حتی در مرحله نمونه سازی-درگیر کنید
- ✓بیشتر جلسات به اشتراک گذاری دانش داخلی را انجام دهید و واژگان و فرایندها را تنظیم کنید.
- ✓پیچیدگی سیستم را دست کم نگیرید و در مورد تمام جزئیاتی که برای تولید محصول لازم است، با کاربر صحبت کنید.

ساختن خط لوله داده‌ها (DATA PIPELINES)

پس از جمع آوری داده‌ها، شما نیاز دارید که نتایج را در کمترین زمان برای کاربران تجزیه و تحلیل، پردازش و خروجی کنید. پردازش داده ها مهمترین بخشی است که شرکت های نوپا باید از آن مراقبت کنند. بنابراین، یک خط لوله داده وظیفه پردازش داده های جمع آوری شده و کمک به دانشمندان داده را برای تجزیه و تحلیل داده‌ها را دارد. خط لوله داده به یک پایگاه داده وصل می‌شود که می تواند یک پلتفرم Hadoop یا یک SQL باشد.

تجزیه و تحلیل کیفیت محصول

تبدیل داده های خام به داده های پخته شده به نحوی که کیفیت محصول تولیدی را به شکلی مختصر و مفید گزارش دهد، یک گام مهم و اساسی برای دانشمندان داده در یک استارتاپ است. این کار می‌تواند توسط ابزارهایی چون تعریف KPI ها، بکارگیری زبان‌های برنامه نویسی قوی در گزارشگیری از داده‌ها مثل R و بکار گیری ETL های مختلف برای استخراج و تبدیل داده‌ها صورت پذیرد.

استارتاپ‌های موفق در حوزه علم داده

مای اسمارت ژن

مای اسمارت ژناستارتاپی در حوزه تحلیل داده‌های ژنتیکی است که ماموریت آن کشف استعدادهای خاص هر فرد است. استارتاپ مای اسمارت ژن توسط منصور حسنی (کارشناسی ارشد هوش مصنوعی) و دکتر دامون نشتاعلی حسنی (دکتری هوش مصنوعی) در بهمن ۹۷ بنیان گذاری شده است.

دیتابین

دیتابین یک سامانه تحلیل رفتار کاربر برای اپلیکیشن‌های موبایلی است. توسعه‌دهندگان می‌توانند SDK دیتابین را در برنامه خود پیاده‌سازی کنند تا داده‌های هر کاربری که برنامه آنها را نصب می‌کند، برای دیتابین ارسال شود. ندا مثنوی (ارشد کارآفرینی) بنیانگذار این استارتاپ است که به همراه مهدی یزدی (کارشناس فناوری اطلاعات) و حامد پالیک (کارشناس کامپیوتر) آن را راه‌انداخته‌اند.

پایشگر هشتگ (سامانه پایش شبکه های اجتماعی)

این محصول به صورت آنلاین و لحظه ای، آخرین اخبار مرتبط با شما را از طریق پایش فضای مجازی در قالب داشبوردی اختصاصی به مدیران و کارشناسان روابط عمومی، بازاریابی، فروش و واحدهای پشتیبانی و ارتباط با مشتریان ارائه می نماید.

استارتاپ NAUTO
دَش‌کمرای هوش مصنوعی برای وسایل نقلیه

دَش‌کمرا نوعی دوربین مخصوص نصب در بخش مقابل راننده یا داخل و شیشه جلوی اتومبیل است که وقتی اتومبیل در حال حرکت است، این دوربین به‌طور پیوسته از راه جلوی اتومبیل تصویربرداری می‌کند. این محصول تحت شبکه، مبتنی بر فناوری ابری است که خطرات تهدیدکَننده‌ی رانندگان را تشخیص می‌دهد. این استارتاپ برای دریافت بازخورد در آخر سفر و آنالیز دلایل تصادفات امکانی فراهم کرده است.

استارتاپ REFLEKTIVE

بازخورد مداوم عملکرد کارمندان

رفلکتیو، نرم‌افزار دریافت و ثبت گزارش و بازخورد عملکرد به کارکنان است. این نرم‌افزار امکانی برای مدیران فراهم می‌کند که از اهداف و شیوه‌ی عملکرد کارکنان چک‌لیستی آنی داشته باشند و بتوانند درباره‌ی هرکدام از آنها، گزارش مستقیمی ارائه دهند. برای کارکنان هم، این نرم‌افزار نوعی کارنامه‌ی درحال‌اجراست که آنها را هر لحظه از نحوه‌ی عملکردشان و میزان رضایت مدیران از عملکردشان مطلع می‌کند.

استارتاپ AYLIEN

پلتفرم هوشمند اخبار

آیلین یک پلتفرم علوم داده در ایرلند است که راه حل‌های تحلیل محتوا با قدرت AI را ارائه می دهد. مدیرعامل این استارتاپ یک ایرانی به نام پارسا غفاری است. ایلین بسته ای از هشت ابزار برای پردازش زبان طبیعی(NLP)، یادگیری ماشین و بازیابی اطلاعات است تا به راحتی معانی ارزشمندی را از اسناد استخراج کند.

دیدن استارتاپ‌هایی بیشتر با موضوع علم داده:

1. European data science start-ups to watch
2. 42 Big Data Startups
3. Deaswatch

منابع:

1. Tansley, S., & Tolle, K. (2009). *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). T. Hey (Ed.). Redmond, WA: Microsoft research.

2. علم داده چیست؟ مجله علمی فرادرس

پایان