1. کلیت مقاله درباره یه سری چالش ها و فرصت های بیگ دیتا در کانتکست پرایوسی هست و اهمیت اون
2. مقاله همون روز اکسپت شده و ۲۰ روز بعد چاپ شده. یعنی طرف به قدری اعتبار داشته تا ایمیل داده گفتن اوکی!
3. اول مقدمه و بعدش یخورده زمین بازی رو بشناسیم یه سری لغت و مهم تر از اون مایل استون های قضیه که دو تا چیز خیلی مهمه. یکی دسته بندی داده و دیگری هم فریم ورک های حریم خصوصی که مهدی برامون میگه
4. هیچی
5. بیگ دیتا رو که همه میدونید از لکچرها هم داشتیم و مقاله اول مقاله اقای گانتنر هم که سه تا وی را مشخصه اصلی بیگ دیتا دونسته. یخورده بیایم و قبل از استفاده از بیگ دیتا و کاربردهای عظیمش به مسئله پرایوسی بپردازیم. یه کاربرد کوچیک پرایوسی در بیگ دیتا از این لحاظ هست که ممکنه چون داده ها distributed هستند و چند تا سرور داریم و ممکنه نخوایم هیچ کدوم از این ها با یکدیگر اطلاعاتشون رو به اشتراک بذارن و یه روش هایی رو پیشنهاد بدیم. بهش میگن federated learning. مقاله دکتر ربیعی درباره پرایوسی که حسین بهش اشاره میکنه یه باب دیگه ای از اهمیت این موضوع رو باز کرده. پس بحث مهمه!
6. تعریف پرایوسی در کمبریج
7. در هر چیزی اول تاریخچه رو نگاه میکنیم. در قرن ۱۴ مثلا مسئله مالیات اهمیت داشته که کسی مطلع نباشه. در قرن ۱۹ با اختراع دوربین یخورده این مفهوم گسترده تر شد. ولی الان با توسعه قوانینی مثل GDPR باهاش سر و کار داریم. اگر یادتون باشه دو سه سال پیش یه قانونی تصویب شد که سایت ها رو ملزم میکرد یه سری کارها انجام بدن. مثلا این که ایسنتاگرام و تلگرام و حتی گوگل رو مجبور کرده کل اطلاعات رو در قالب یه زیپ تحویل بدن و حتی در سال ۲۰۱۶ یه قانونی به نام حق فراموش شدن وجود داره. یه دادگاه اروپایی اومد تصویب کرد موتورهای جستجو در صورت درخواست باید «دقت کنید باید» اسم هر کسی رو حذف کنن. دادگاه های اروپا هم که میدونین اگر یکیشون یه چیزی تصویب کرد کلا برای بقیه سند هست و بقیه میتونن به این استناد کنن.
8. خب مسئله پرایوسی در یک دیتاست رو نمیشه با حذف ستون های هویتی اوکی کرد؟! جوابش اینه نه! دو تا تحقیق انجام شده روی دو تا دیتاستی که هیچ اطلاعات هویتی توشون نبوده. مثلا یکی دیتاست ۱۵ ماهه ۱.۵ میلیون نفر را داشته و تونستن با یه سری پردازش با دقت ۹۵ درصد هر فردی را دقیقا با ۴ تا مکان و زمان مشخص که بهش میگن temporal spetial مشخص کنن. دقیق. یه تحقیق دیگه اومده با ۳ میلیون رکورد تراکنش ۱.۱ میلیون نفر با دقت ۹۰ درصد با ۳ تا نقطه دقیق مشخص کنه هر فردی رو بصورت یونیک. پس باید در معنای id یخورده تامل کنیم و دیگه شاید ایدی فقط اون لیبلی که از بچگی روی ما زدن نباشه. با یه سری فیچر میشه به سادگی هر فردی رو uniquely identify کرد.
9. دو تا دسته بندی داریم: مثلا در اولی هکر در مورد یک نفر یک اطلاعاتی داره مثلا میدونه من بعداز ظهر ها میرم شیر بخرم از فلان مغازه. بعد با تراکنش های دیتاست های پابلیک میتونه من رو شناسایی کنه. این یه دسته بندیه. دسته بندی دوم هم درباره تعامل ما و نحوه ارتباط ما هست که ممکنه یه سری ویژگی ها رو بفهمه مثلا وقتی من از ترافیک ssl استفاده میکنم و از نحوه ارتباط گرفتن من و پروتکل هام من رو تشخیص بدن که واقعا دمش گرم کسی که بتونه اینکارو بکنه!
10. درباره حفاظت از پرایوسی و روش های اون علوم زیادی توسعه پیدا کرده اند که مثلا رمزنگاری علوم ارتباط و حتی تئوری اطلاعات رو داریم.
11. مثلا تو بحث رمز نگاری این اقا مطرح میکنه که بزرگترین چالش اینجا در اسکیل بزرگ بیگ داده گوشی های یخورده قدیمی تر نمیتونن الگوریتم های رمزنگاری یخورده قوی تر رو اجرا کنن و ممکنه مشکلات پیش بیاد.
12. در روش های ارتباطی پروتکل های اوپن سورس داریم. اوپن سورس بودن پروتکل خیلی مهمه. تو کتاب تننباوم که قدیم رفرنس شبکه بود. فصل اخرش در مورد امنیت تقریبا مهم ترین نکته در مورد یه پروتکل ارتباطی رو شفاف بودنش میدونست. چون مشکلاتش کم کم به طول زمان محققان امنیتی کشف میکنن. همینه که لینوکس الان by far جلوتر ویندوزه! سیستم های تور و شبکه پیازی و اینها رو هم داریم که اگر از من بپرسید شاهکاره به نظرم. ترکیب پروتکل های رمزنگاری متقارن نامتقارن و کلید و خیلی جالبه ایده شون.
13. بخش دوم مقاله این اقا یک سری ترم رو تعریف کرده.
14. در شکل مشخصه. تولید کننده داده اون کسی هست که دیتا خام تولید میکنه و به نوعی دیتاش ذخیره میشه. حالا ممکنه پست کنه ممکنه در دیتابیس ذخیره بشه دیتاش. ممکنه اکتیو باشه ممکنه پسیو باشه. دیتا کیورتور اون شخص یا سازمانیه که دیتا رو نگهداری و احتمالا توزیع میکنه. دیتایوزر به یک معنا اونی هست که استفاده میکنه از دیتا. مثلا ما و اتکر هم که مشخصه. دیتا یوزر شاید یخورده گول زننده باشه اسمش. و سه تا عملیات خیلی مهم در اینجا مشخص شدن. یکی جمع اوری دیتاهست و بعدش باید دیتا ایدی و بقیه چیزهای مشخص ازش حذف بشن و بحث سوم ارتباط هست.
15. درباره جمع آوری دیتا اقای اندروانجی رو من خیلی وقته تو لینکداین دنبال میکنم پستی چیزی میذاره میخونم ببینم چطوریه. خیلی ادم جالبیه. حرفش اینه که ما باید از سیستم های بر مبنای مدل به سیستم های بر مبنای دیتا بریم. یعنی چی یعنی اقا تا الان ما مدل ها را تیون میکردیم که تیون کردن پارامترها در شان ما نیست. باید به سمت روش های سیستماتیکی بریم که دیتای خوب تولید کنیم ولیبل هایی که میدیم درست باشن. اینجوری با مدل های خیلی شاید ساده هم دقت خیلی بالا میگیریم.
16. یه دسته بندی ای هم بر روی فیچرها میده و میگه فیچر ها بعضا مثل ایدی مشخص میشه دقیقا فرد کیه. با بعضی فیچرهای quasi میشه تشخیص داد فرد کیه. یه سری فیچرها اطلاعات حساس هستند و بقیه رو هم Other دسته بندی کرده البته به نظرم این دسته بندی جامع و مانع نیست. خیلی جالب دسته بندی نشده.
17. یه گریزی هم به کاربرد ها بزنم و یه کاربرد رو معرفی کنم و بدم دست مهدی. در نظریه بازی که درسش هم ارائه میشه یک سری مفاهیم جالب داریم که میتونه اینجا کمک کنه. اونجا حرفمون اینه که اقا میتونیم خیلی وقت ها مکانیزم هایی رو دیزاین کنیم که افراد به نفعشون باشه که اونجوری که ما میخوایم بازی کنن. حالا تو بازار کاربرد داره تو اقتصاد. یه چیز جالبش حراج های قیمت دوم هست که نقطه تعادل نشش معادل حراج قیمت اوله ولی نشون داده شده بهتره!