



ISC 互联网安全大会



360 互联网安全中心



大数据的安全、融合共享与运用

冯景辉

百度安全技术总监

2018 ISC 互联网安全大会 中国·北京
Internet Security Conference 2018 Beijing·China
(原中国互联网安全大会)

目录

大数据安全

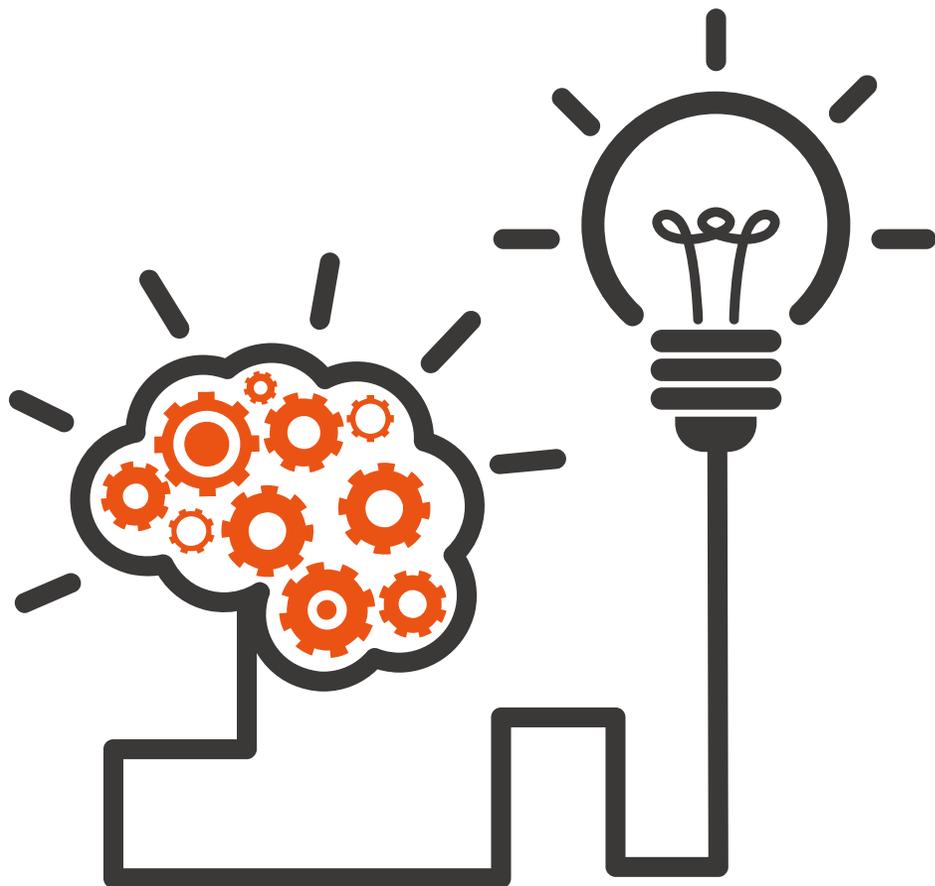
狭义数据安全与广义数据安全

再谈数据泄漏

数据流通的多个环节

数据共享与脱敏

数据安全的狭义与广义



核心数据库防护是第一道考题



弱密码

有多少拖库是弱密码引起的



数据库公网访问

建议的服务器部署模式



源码泄漏

Shadow IT 引发的风险



无分级、无风控

认证鉴权，PII 信息比对等内容应该以服务的形式存在，而不是直接访问数据库

数据流通的多个环节



ISC互联网安全大会



360互联网安全中心



数据共享与脱敏

删除PII信息就够了吗？



ISC 互联网安全大会



360 互联网安全中心



2006年，美国最大的影视公司之一 Netflix，举办了一个预测算法的比赛（Netflix Prize），比赛要求在公开数据上推测用户的电影评分。Netflix 把数据中唯一识别用户的信息抹去，认为这样就能保证用户的隐私。但是在 2007 年来自 The University of Texas at Austin 的两位研究人员表示通过关联 Netflix 公开的数据和 IMDb（互联网电影数据库）网站上公开的纪录就能够识别出匿名后用户的身份。三年后，在 2010 年，Netflix 最后因为隐私原因宣布停止这项比赛，并因此受到高额罚款，赔偿金额总计九百万美元。

01 PII

个人身份标示数据，例如电话号码，身份证号，银行卡号，账号等。在进入分析环节这部分数据一般直接删除。

02 PI

私人数据，虽然不能标示个体，但属于分析个体时重要属性，例如性别、年龄、学历

03 Sensitive attributes

数据分析的主体，是数据画像的结果，例如购买偏好，收藏历史等。这部分数据在数据分析时一般不能删除。并且会作为结果展示。

匿名化



ISC互联网安全大会



360互联网安全中心

	姓名	性别	年龄	手机品牌	购买偏好
	Kitty	女	20-25	iphone	化妆品
	Kath	女	20-25	oppo	厨具
	Bob	男	25-30	vivo	家电
Sam	Abe	男	35-40	iphone	电子产品
	Frank	男	35-40	iphone	电子产品
	Sam	男	29-30	vivo	图书
	Lavinia	女	35-40	vivo	图书
	Sammi	女	35-40	oppo	日用品

K匿名化

K-匿名化的目的是保证公开的数据中包含的个人信息至少 $k-1$ 条不能通过其他个人信息确定出来。也就是公开数据中的任意 PI 信息, 相同的组合都需要出现至少 k 次。

差分

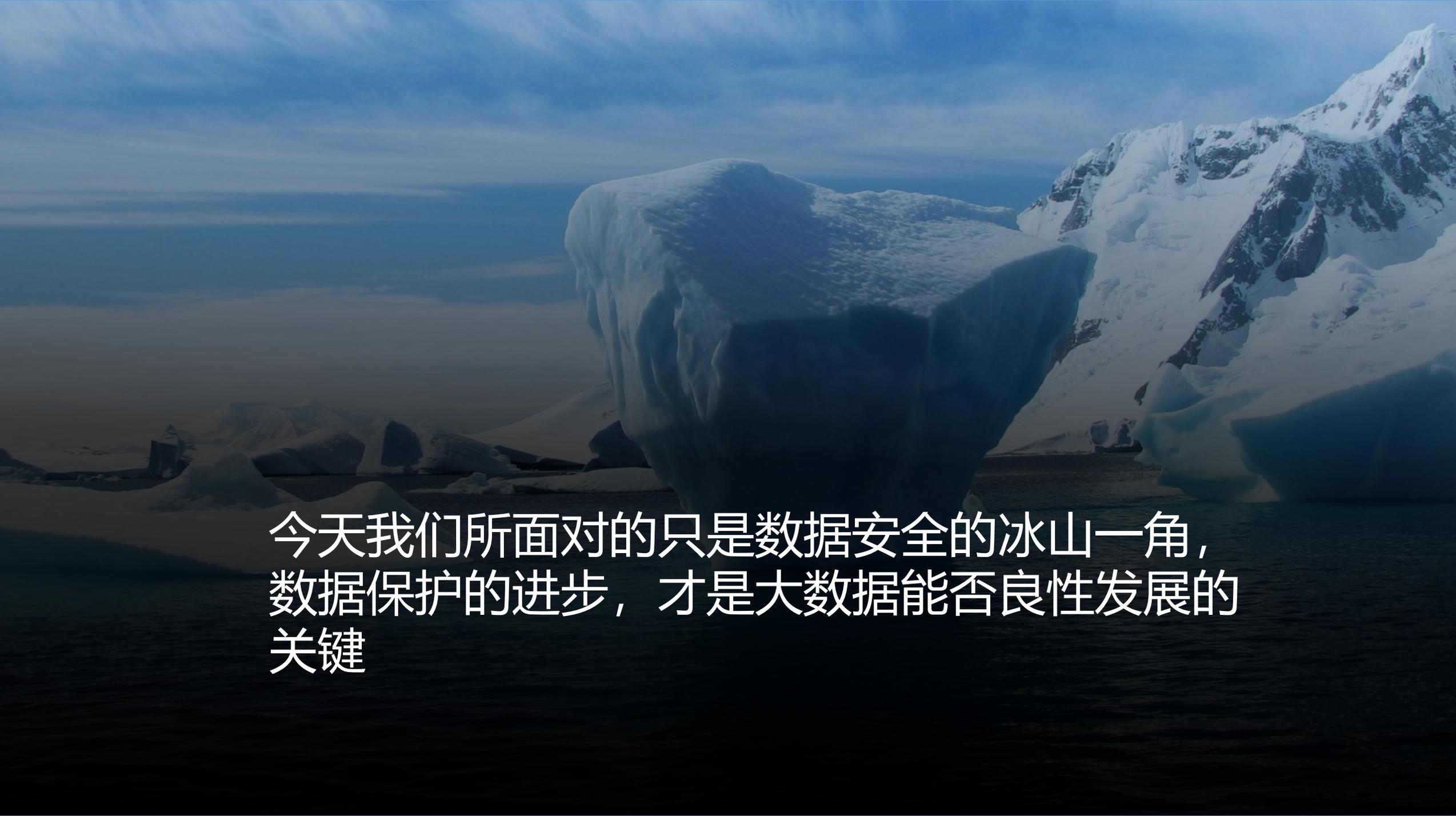


数据噪音

中心化与本地化数据
差分



苹果在Iphone上使用的
本地数据差分算法

A large iceberg floats in the ocean, with a massive, dark, and jagged portion submerged below the surface. The visible part above water is much smaller and smoother. In the background, a snow-capped mountain rises against a blue sky with light clouds. The overall scene is dimly lit, with a blue and grey color palette.

今天我们所面对的只是数据安全的冰山一角，
数据保护的进步，才是大数据能否良性发展的
关键



ISC 互联网安全大会



360 互联网安全中心

谢谢!

2018 ISC 互联网安全大会 中国·北京
Internet Security Conference 2018 Beijing·China
(原中国互联网安全大会)