

烽火科技的大数据之路

烽火通信科技股份有限公司

安全产品线 大数据首席专家：王胤然

写在最前面的话

简介公烽火的基本情况和从事的行业信息

公司简介

烽火科技集团（FiberHome Technologies Group）是中国优秀的信息通信领域产品、解决方案与综合服务提供商，“中国光谷”的核心企业，直属国务院国有资产监督管理委员会管理。1974年正式成立，经过三十年的发展，集团目前已形成覆盖光纤通信技术、数据通信技术与无线通信技术三大产业的发展格局，旗下拥有多家子公司。烽火科技集团的前身——武汉邮电科学研究院

（WRI）是中国光通信的发源地，也是国内唯一集光通信三大领域：通信系统、光纤光缆、光电器件的研究、开发、生产和销售于一体的科研与产业实体。是中国电子信息百强和软件百强企业。经国家批准为：“国家光纤通信技术工程研究中心”、“国家光电子工艺中心”、“国家高新技术研究发展计划成果产业化基地”、“亚太电信联盟培训中心”，06年被国家确定为103家创新型试点企业

然而我们不止卖光纤

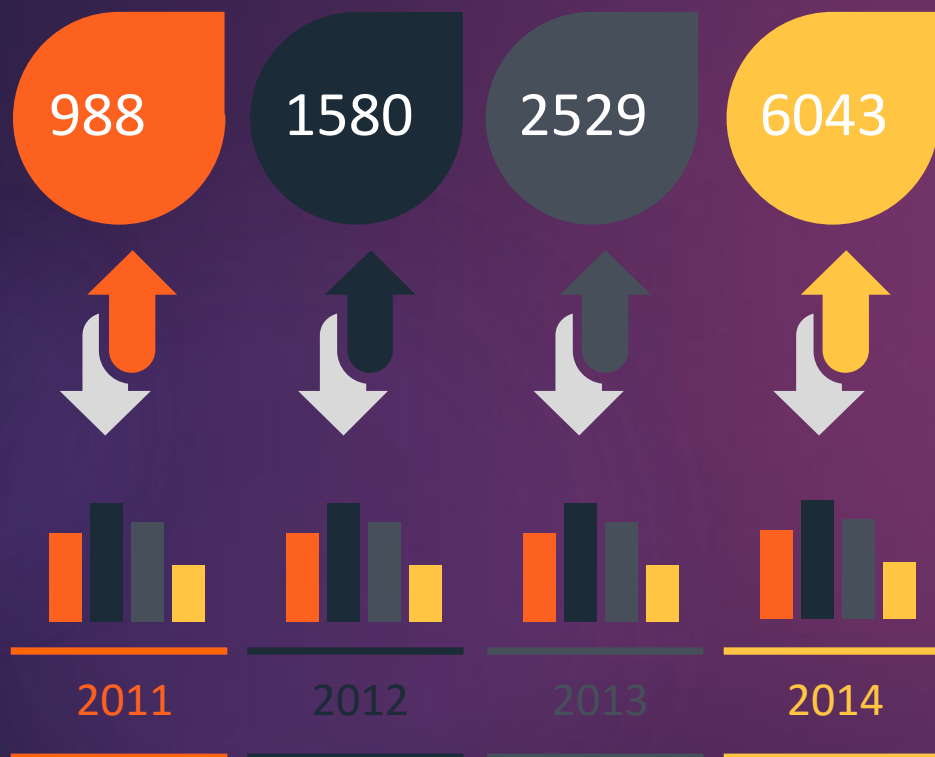


南京烽火是国内拥有完全网络信息安全资质的三家企业之一，凭借其电信级别的数据采集信息提取能力和海量PB级别大数据分析处理能力坐拥行业信息安全领域70%以上的市场份额。

理论基础

从数据的发展规模以及面对的实际处理需求出发建立行业大数据处理方法论

老生常谈的大数据飞速发展



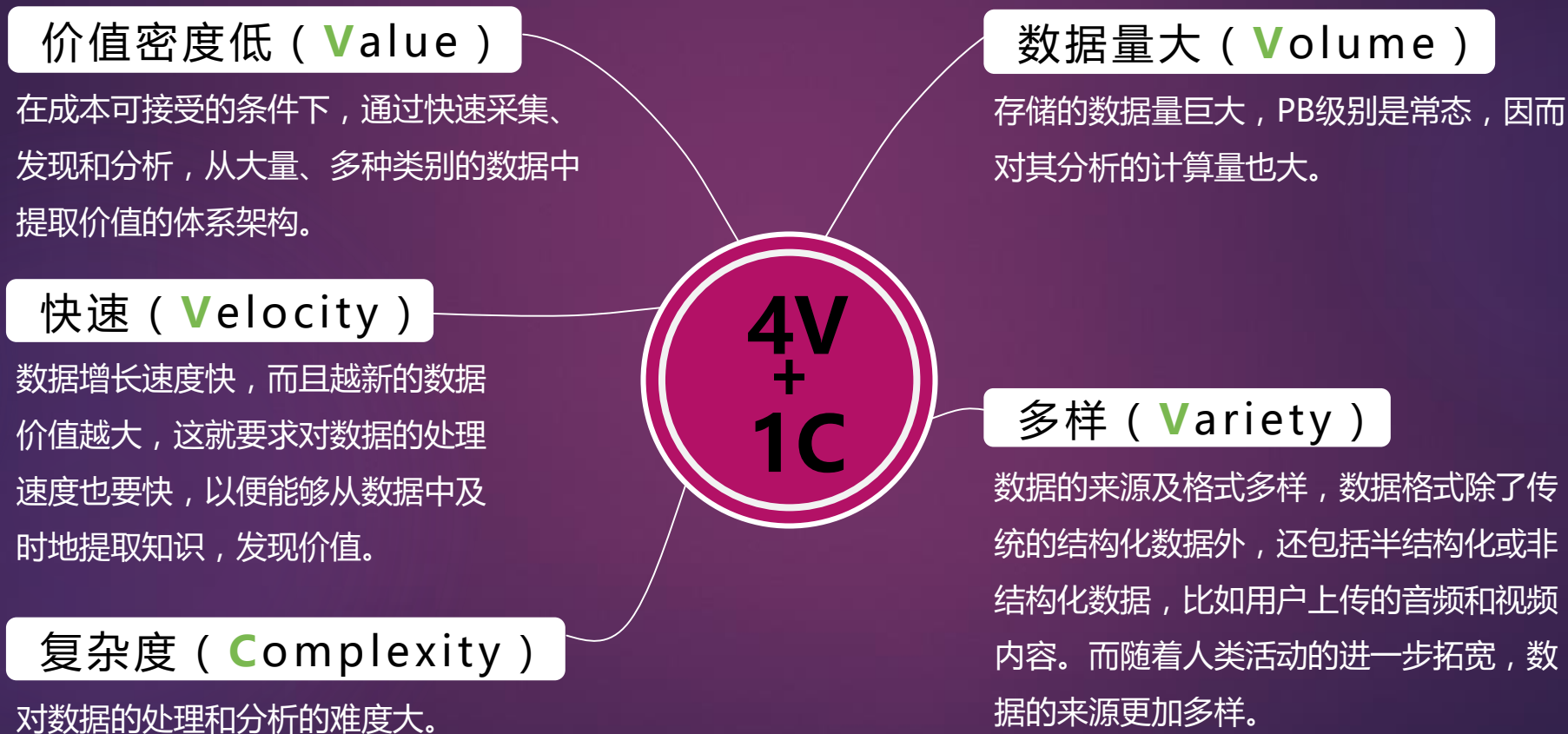
数据增量大

信息充斥着我们的生活，信息改变着我们的生活，社会的每个个体无时无刻产生着各式各样的数据，信息量巨大。随着信息化程度越来越高，数据增量越来越大。

数据增量变化快

数据本身体积大，产生速度快，随着硬件产品、网络技术的不断更新和发展，数据可被利用和存储的量越来越大，导致数据增量越来越大

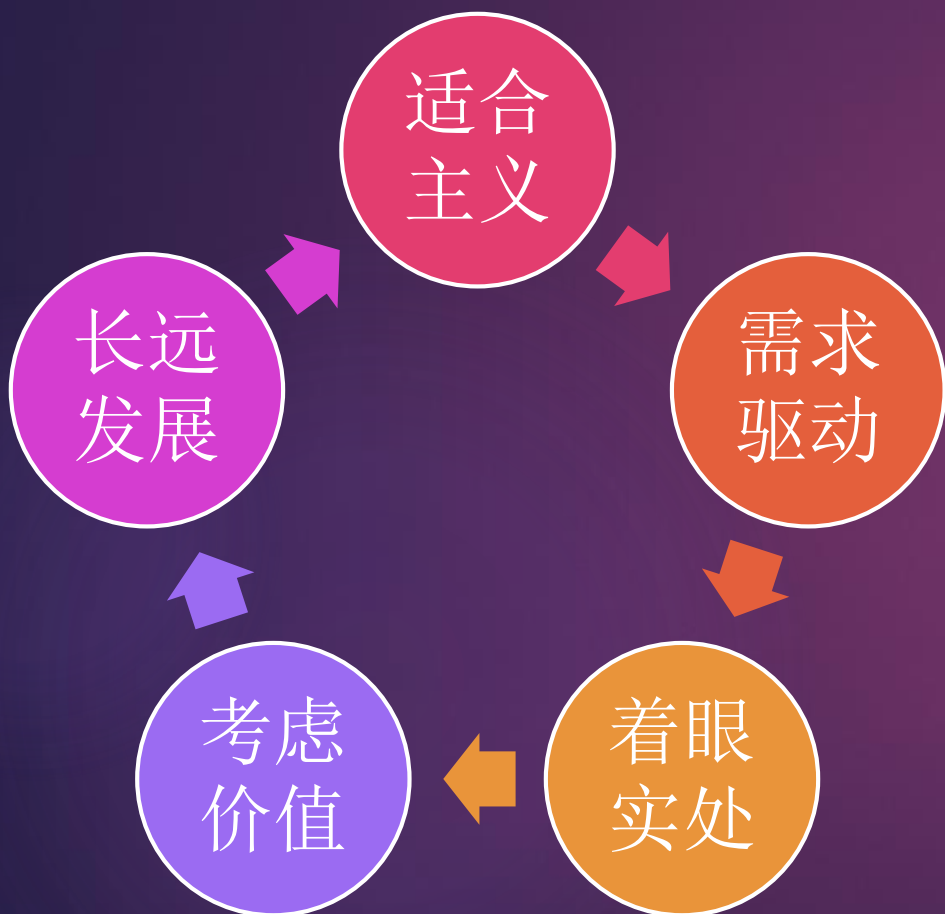
老生常谈的大数据飞速发展



大数据的三宗罪

- ▶ 大数据“神话主义”言论
 - ▶ 利用大数据可以上知天文下知地理！
 - ▶ 利用大数据可以了解过去“预知”未来！
- ▶ 大数据包治百病
 - ▶ 传统IT架构出现问题时，能拯救你的只有大数据技术
 - ▶ 大数据可以提高你系统效率，10倍，100倍甚至1000倍！
- ▶ 大数据是趋势，任何的系统架构都要向大数据架构上转型
 - ▶ “大数据时代”来临，IT信息发展需要顺从趋势大力发展大数据技术，将传统架构逐步向大数据产品架构转型！

大数据时代的几点忠告



现在的IT架构是否已经满足需要？是否简单的扩展就或者调整就可以满足要求？

回归最基础的需求分析阶段，原始需求是什么？需要达到什么样的目标？

现有基础是什么？如若引入大数据技术，需要多大的投入？达到的效果是否可以满足需求

引入大数据技术，给系统究竟带来了多大的价值，与之前的投入是否成正比

引入大数据技术后是否可以满足未来几年的数据规模或者需求发展需要

信息安全行业的数据特点

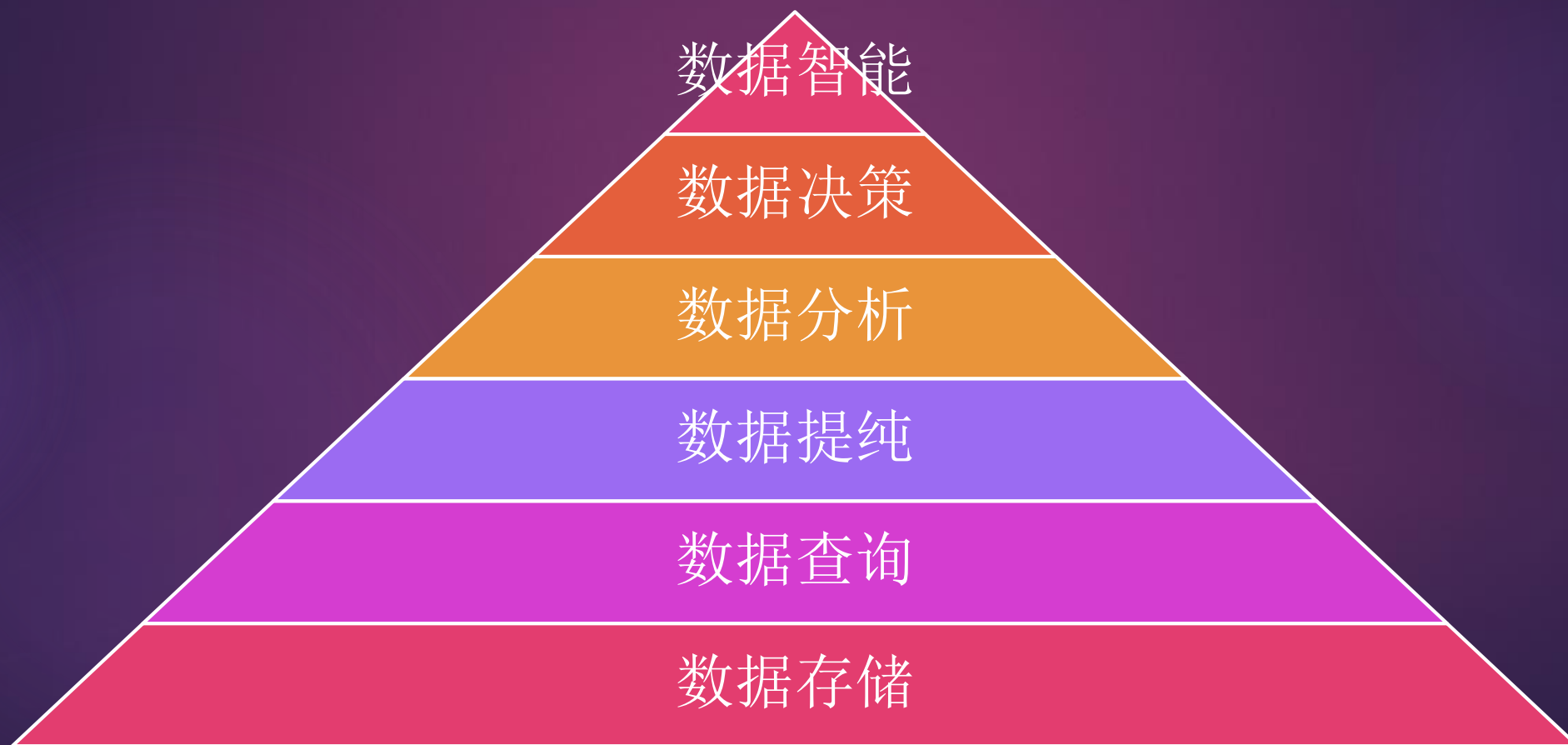
烽火面对的大数据日常

- ▶ 每天处理的网络数据达到685亿条，单日存储量高达170TB。
- ▶ 行业特性决定的三级建设模式，涉及集群100余套，跨越全国28个省和直辖市、100余个地市。
- ▶ 三级联动，交换汇聚互相调用。集群规模从几台到几百台不等。

信息安全行业的数据特点

	互联网	信息安全行业
数据存储周期	数据长期存储，有的数据存储达到2年以上，每天新接入数据相对历史数据非常少。	数据存储周期很短，大多是几个月，新接入数据相对历史存储数据非常大。
集群规模	数据不停累计，集群规模越来越大，计算资源非常丰富。	数据基本很少累计，集群规模固定，计算资源非常缺乏。
查询	互联网查询条件相对固定，并且存在大量的热点数据，可以通过cache、CDN技术解决查询速度	查询条件非常灵活，并且不存在热点数据，只能通过查询数据库获取结果。大量的随机IO使的数据库压力非常大。
计算任务	集群规模庞大，计算资源丰富，单个磁盘涉及的数据非常少，IO压力较小。	集群规模有限，计算资源匮乏，当个磁盘涉及的数据量较大，IO压力较大。
计算模型	很多业务模型和自身计算能力相关，设计业务时尽量让计算只涉及当前数据。减轻计算和IO压力	很多业务已经形成标准，计算中涉及到大量的历史数据操作。需要的计算资源和IO压力较大

信息安全行业大数据发展的解决之道



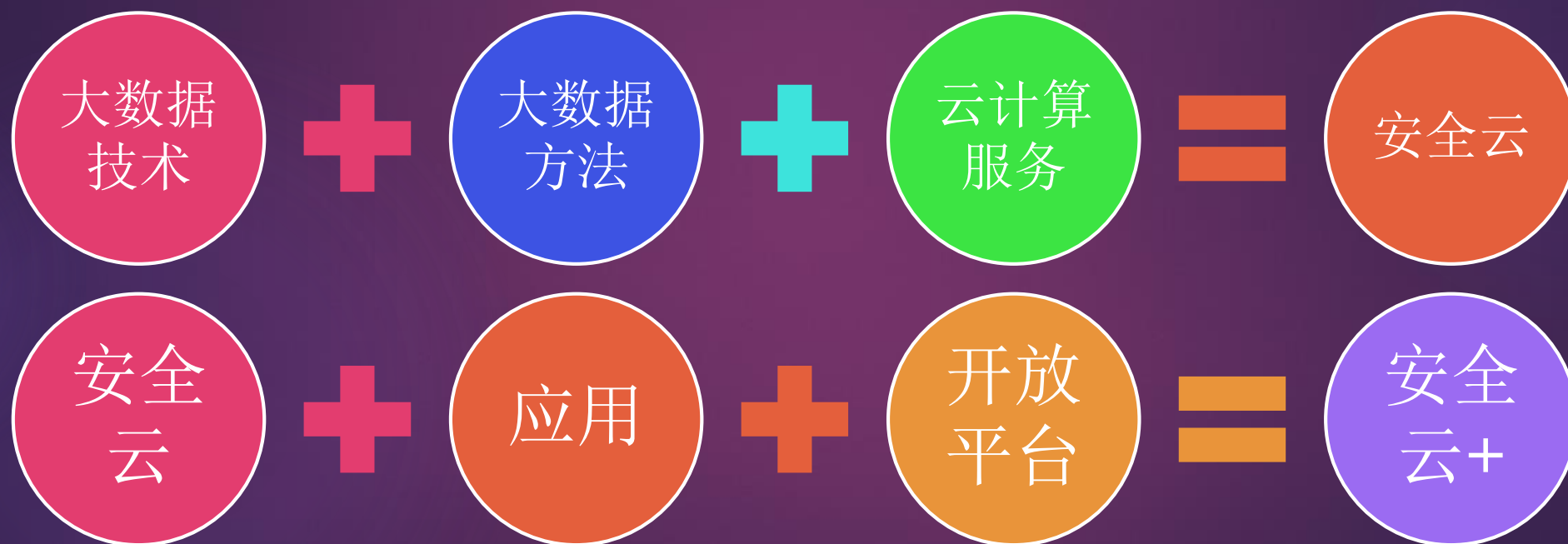
信息安全行业大数据到底怎么玩儿？

一杯美酒的故事-揭秘大数据处理的方法



数据相当于粮食，经过转换会产生新的价值！
数据处理的过程也是同样的道理！

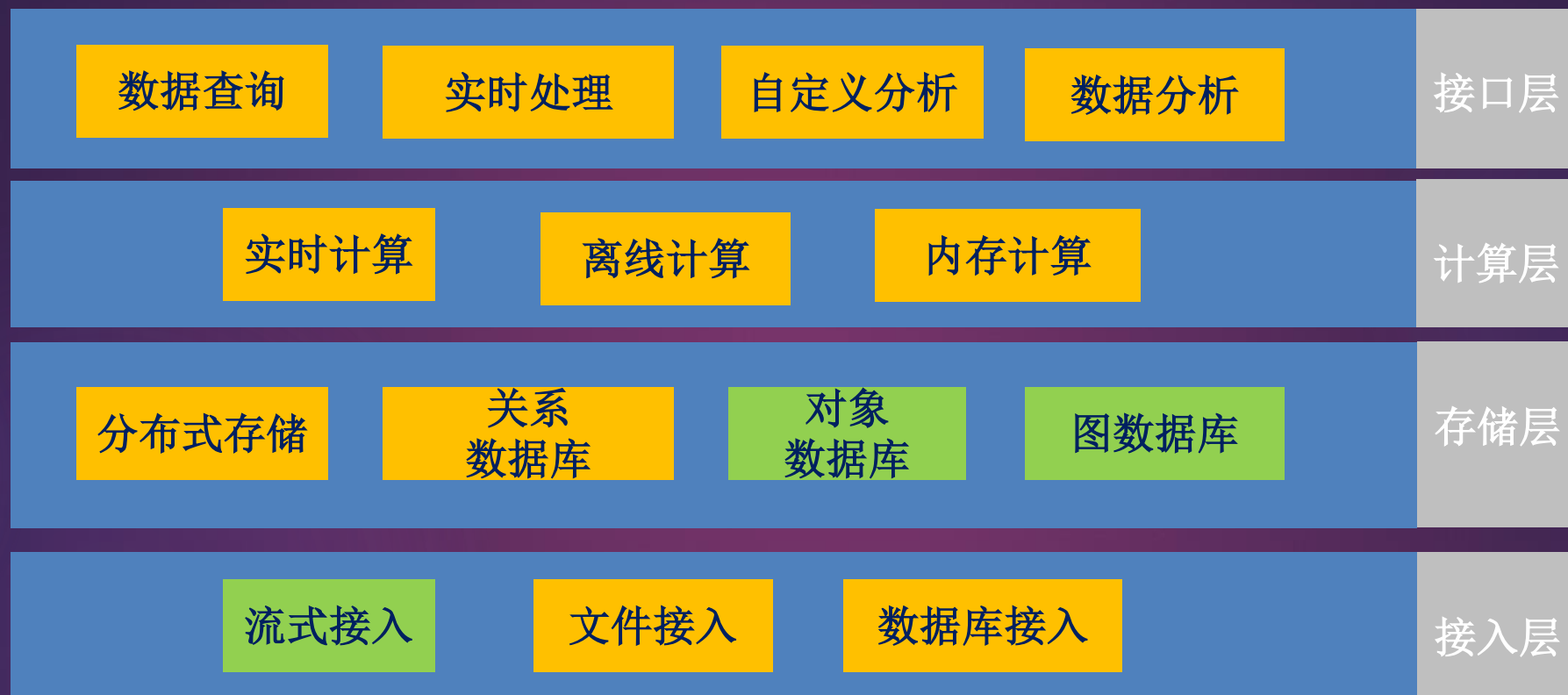
安全行业大数据处理的方法论



架构之路

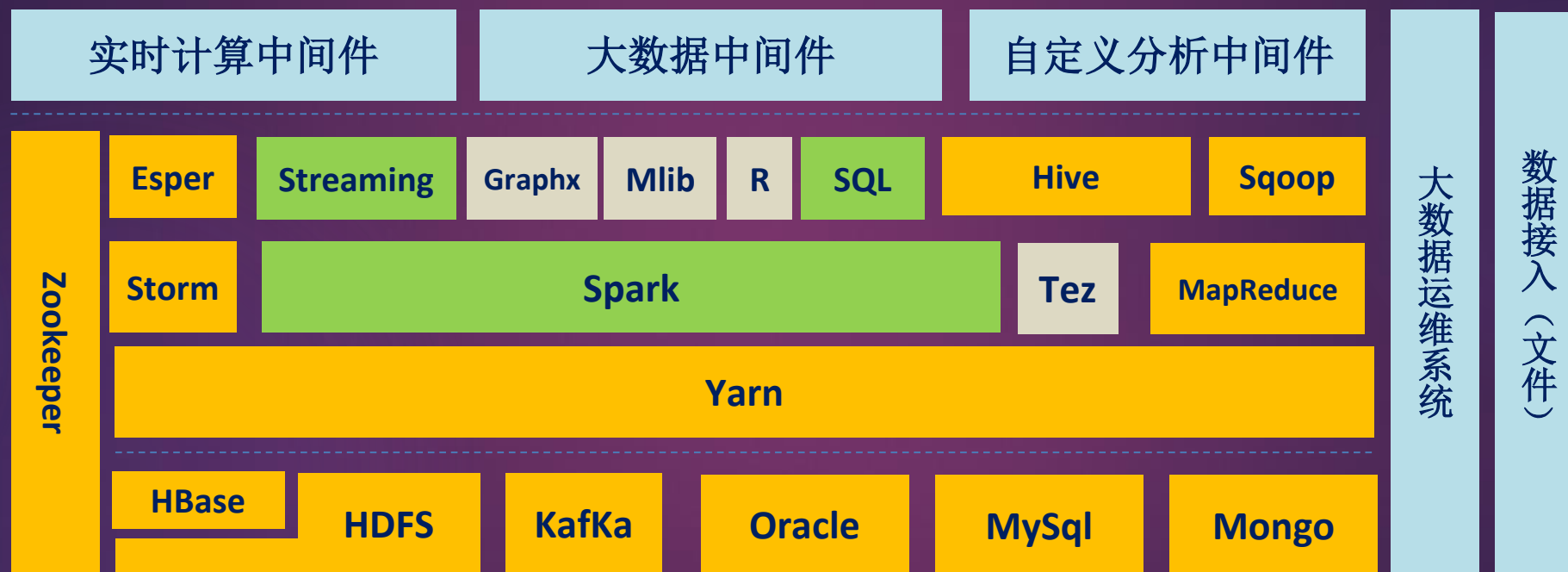
路漫漫其修远兮，踏踏实实做设计。面对需求，信息安全行业烽火大数据平台的基础架构

需求下的产品架构设计



平台提供大数据端到端的解决方案，包括数据接入、存储和计算，融合实时计算、离线计算和内存计算于一体；对外提供统一的访问接口，屏蔽平台复杂性；

烽火安全大数据平台的基础架构



通过将开源技术和自主研发相结合的方式构架自己的大数据平台，进行了针对性的扩展和调优。

手段之路

分享两个烽火成熟的大数据关键技术解决方案

FH-HBASE二级索引方案

Query



Index

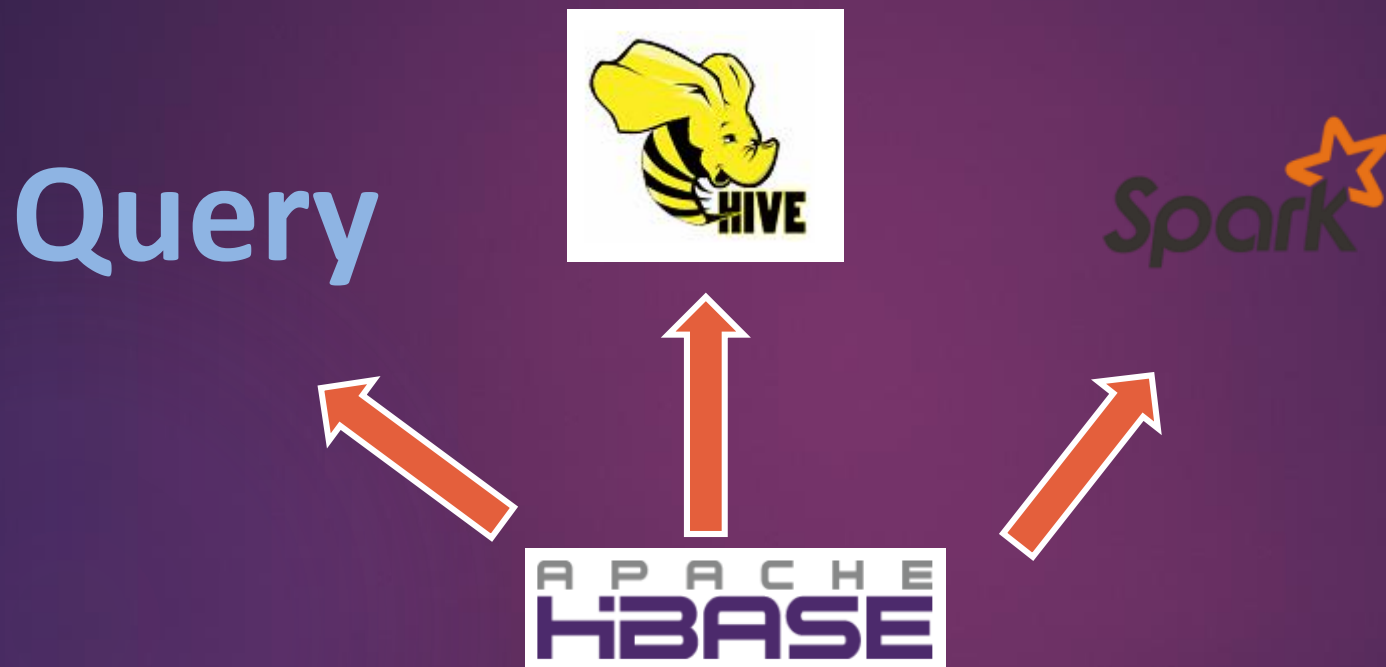


- 基于HBase构建全局索引
- 存储和性能比较平衡
- 查询性能可横向扩展
- 故障自动接管，对应用完全透明

FH-HBASE二级索引方案

		D	F		P
		a	f	drk	f
数据区	D xxx	c1 c2 ...			
索引区	F xxx		f1 f2 ..	D xxx	f1 f2 ..
	G xxx		f1 f2 ..	D xxx	f1 f2 ..
	... xxx		f1 f2 ..	D xxx	f1 f2 ..
	行rowkey	数据列族	过滤列	数据区RowKey	过滤列

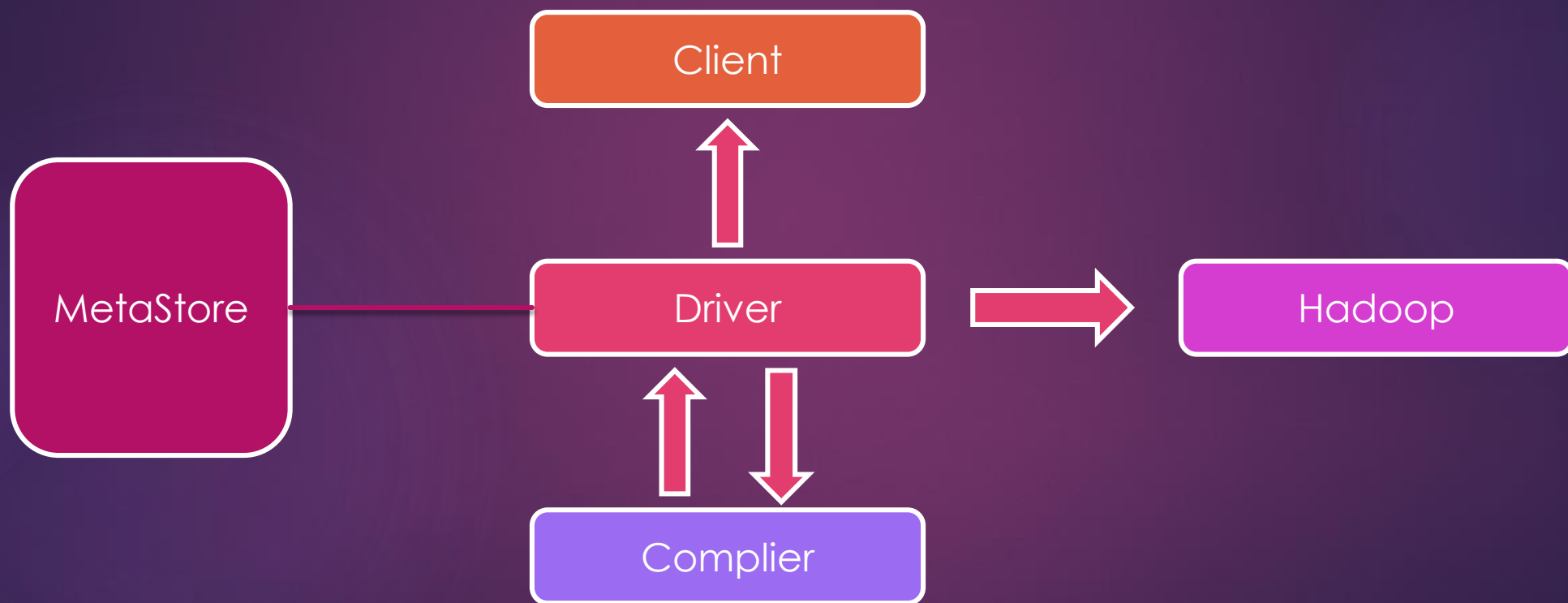
FH-HBASE二级索引方案 分析平台的支撑



- 同一套存储兼顾查询和分析，减少数据存储
- 高速并行的数据扫描
- 分析和查询条件自动下推，减少IO消耗

FH-HBASE二级索引方案

以兼容HIVE为例



FH-HBASE二级索引方案

以兼容HIVE为例

Paser

将HQL转换抽象成语法树

Semantic Anlayzer

将抽象语法树转换为查询块

Logic Plan Generator

将查询块转换为逻辑查询计划

LogicalO pitimizer

重写逻辑查询计划

Physical Plan Generator

将逻辑计划专程物理计划（MR）

PhysicalO ptimizer

选择最佳执行策略

AST

QB

OP TREE

OP TREE

TASK
TREE

TASK
TREE

注入私有数据表结构

注入私有数据表索引结构

转换为索引Task



烽火大数据中间件技术



ZOOKEEPER

大数据中心QueryService

HBASE

大数据中心TaskService

MAPREDUCE

HDFS

任务服务TaskService

提供了MR任务的调度和执行

查询服务QueryService

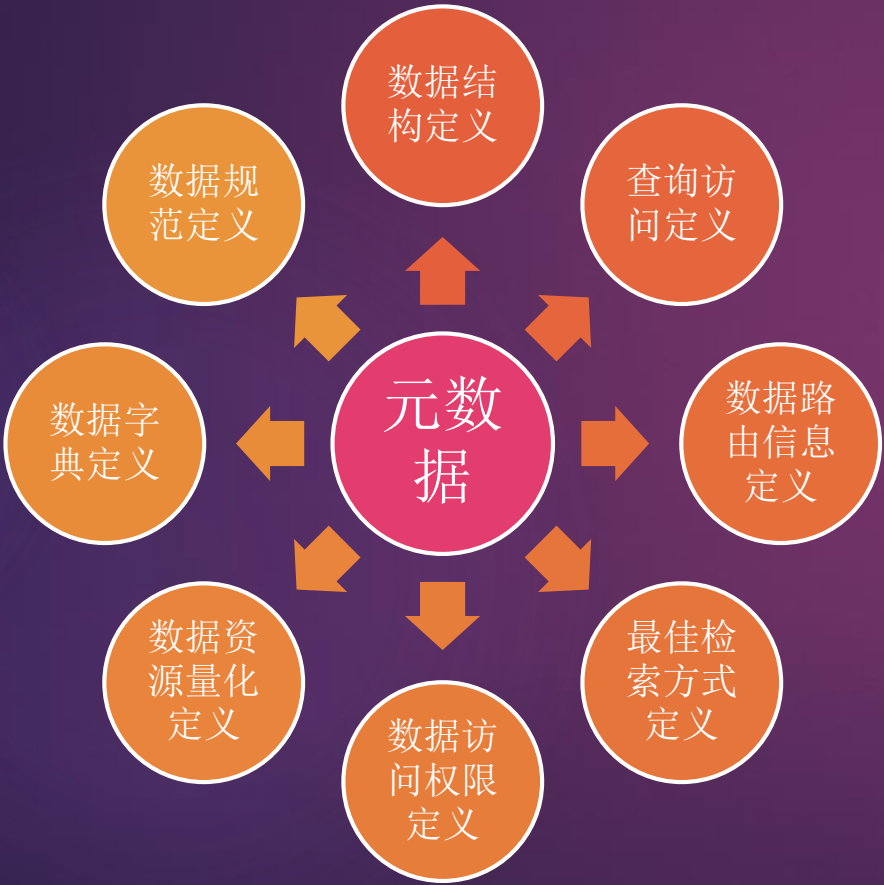
提供了HBASE 的数据查询功能



发展方向

在即将到来的2016年，我们计划推进的大数据相关解决方案

数据资产化的梳理与管理



通过元数据盘点数据资产

- 数据种类盘点
- 数据量盘点
- 数据源盘点
- 数据索引盘点

通过元数据体现数据资产价值

- 数据访问权限控制
- 数据索引访问路由控制
- 数据规范的定义

谢谢！

王胤然 MAIL: MARKSWANG@126.COM WX: MARKSWANG

2015年12月