HUAWEI

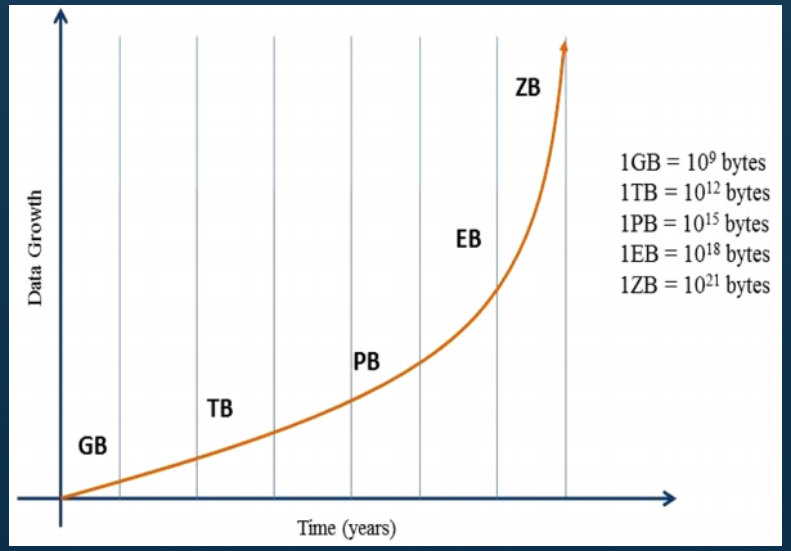
Day17 轻松探索数据背后的价值
——数据湖探索

目录

大数据时代

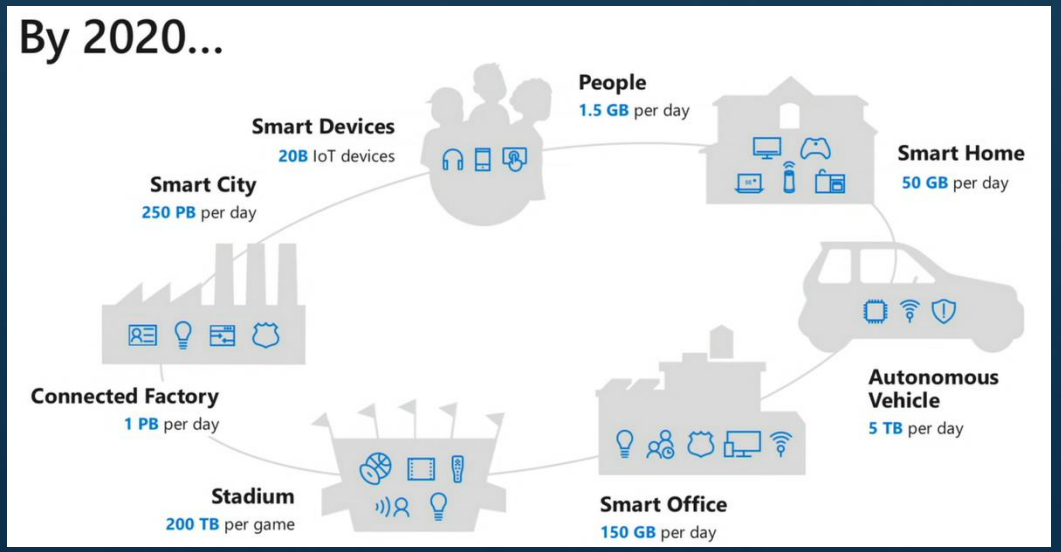
当前社会每年产生的数据呈指数级增长，数据分析已经成为了企业的经营管理者们极为重视的一项活动内容。

据《福布斯》2017年统计，已有53%的企业正在采用大数据分析。分析机构 Gartner 在2017年对 2500 名 CIO 实施调查后发现，所有类型的企业都把“商业智能和数据分析”的支出列为首要的投资事项。



数据量增长趋势图

source:Int. J. Bioinformatics Research and Applications, Vol. 14, Nos. 1/2, 2018



微软预测2020年各行业的数据量

source:State of the Art Big Data as a Service,insight.com

HUAWEI

大数据的特征

不同于普通的数据，大数据的复杂特征要求需要经过专业复杂的处理，才能最终获得有价值信息。

大数据的应用示例

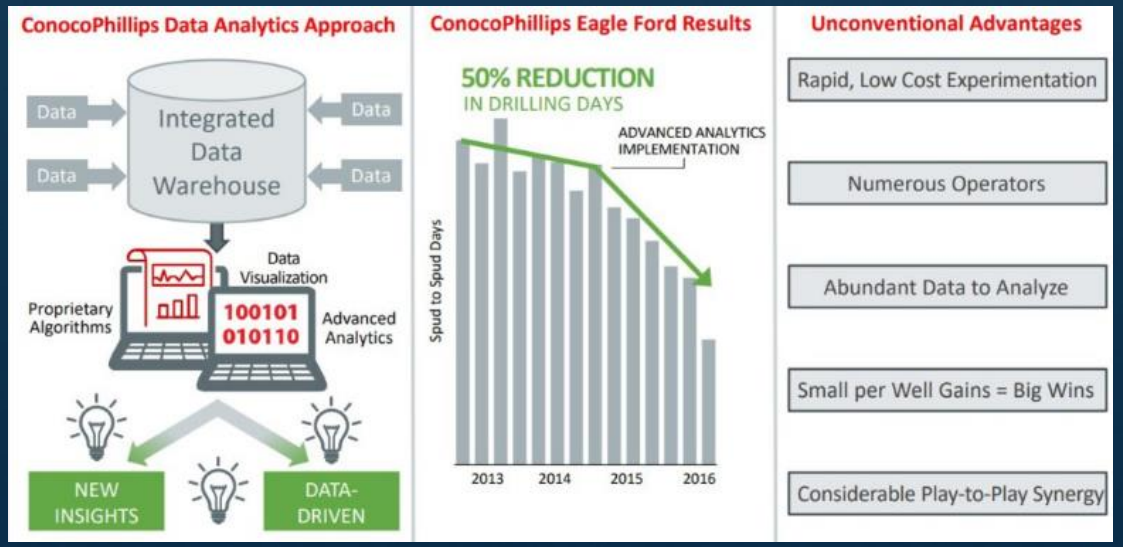
- Google成功预测冬季流感

2009年，Google通过分析5000万条美国人最频繁检索的词汇，将之和美国疾病中心在2003年到2008年间季节性流感传播时期的数据进行比较，并建立一个特定的数学模型。

最终Google成功预测了2009冬季流感的传播，甚至可以具体到特定的地区和州。

- 康菲石油公司（ConocoPhilips）的综合数据中心（IDW）项目

康菲公司的IDW项目，运用大数据分析帮助公司做出更好的决策，实现低投入高收益的目标。



IDW项目帮助康菲公司降低成本、提升效率

目录

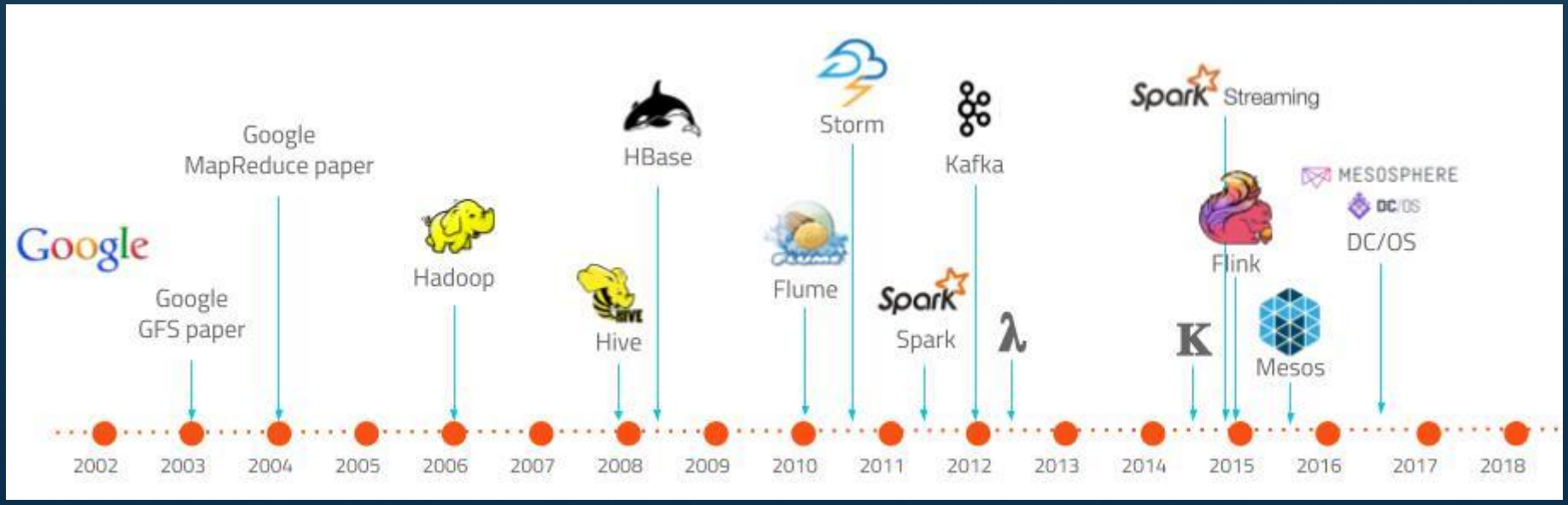
大数据技术

大数据技术主要包括：数据采集、数据存储、数据计算和分析、数据的呈现与应用。其中数据的计算和分析技术是大数据技术的基础和核心。

Source:Big Data: Moving from Technology to Business Value Delivery, informatica.com

大数据技术的发展历程

大数据概念在1998年由SGI首席科学家John Masey在USENIX大会上的一篇论文中首次提出，使用了大数据来描述数据爆炸的现象。但大数据技术的开端是2003-2006年Google发布的GFS、MapReduce和BigTable三篇论文。基于此，产生了开源的Hadoop，开启了大数据技术和应用飞速发展的大门。



大数据技术年表

source:From Lambda to Kappa: evolution of Big Data architectures, paradigmadigital.com

大数据的批处理和流处理

批处理
Batch Processing

批处理主要操作大容量静态数据集，并在计算过程完成后返回结果。

适用数据特征：

- 有界：批处理数据集代表数据的有限集合
- 持久：数据通常始终存储在某种类型的持久存储位置中
- 大量：批处理操作通常是处理极为海量数据集的唯一方法

处理框架：
Hadoop/Spark/Flink

流处理
Stream Processing

和批处理处理静态数据不同，流处理系统是对随时进入系统的数据进行计算。

数据特征：

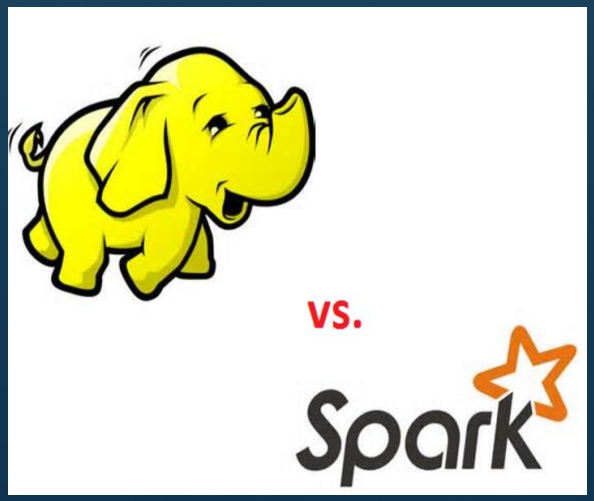
- 无穷：源源不断有新的数据产生
- 实时产生：数据随时产生，要求计算低时延、可持续

处理框架：
Storm/Samza/Flink/Spark streaming

大数据技术框架和系统

Hadoop 和 Spark

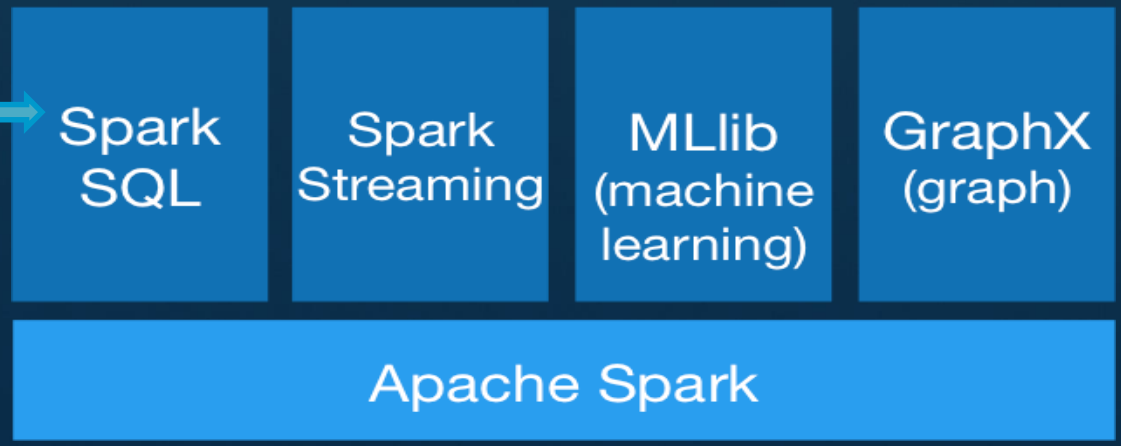
计算是大数据技术的核心，Hadoop和Spark是当前最流行两大计算框架。由于在计算速度及生态上的优点，使用Spark作为计算引擎的企业或机构也越来越多。



对比项	Hadoop	Spark
用途	分布式存储+分布式计算	分布式计算
计算框架	MapReduce模型	通用计算方法
数据存储	磁盘（HDFS）	磁盘或内存
迭代计算	不支持	支持
处理方式	批处理	批处理+流处理
计算效率	-	相比Hadoop，磁盘上运算快2~10倍，内存上快100倍

Spark还具有以下优势：

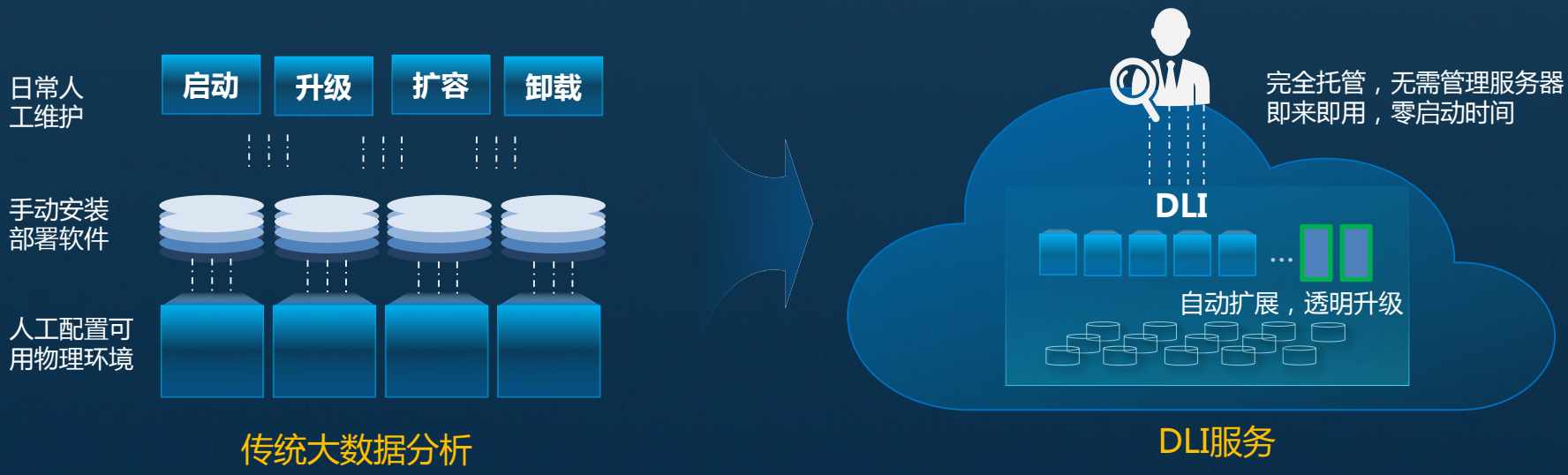
- 丰富的生态 SQL,streaming,Mllib…
- 支持多种API Python,Java,scala…
- 支持多种部署平台 Mesos,yarn,k8s…



目录

华为云数据湖探索服务

华为云数据湖探索（Data Lake Insight, DLI）服务是完全托管的大数据处理分析服务。用户不需要管理任何服务器，即开即用。支持标准SQL，兼容SparkSQL，支持多种接入方式，并兼容主流数据格式。数据无需复杂的抽取、转换、加载，使用SQL或Spark程序就可以对华为云上CloudTable、RDS、DWS等异构数据进行处理和分析。



https://www.huaweicloud.com/product/dli.html

HUAWEI

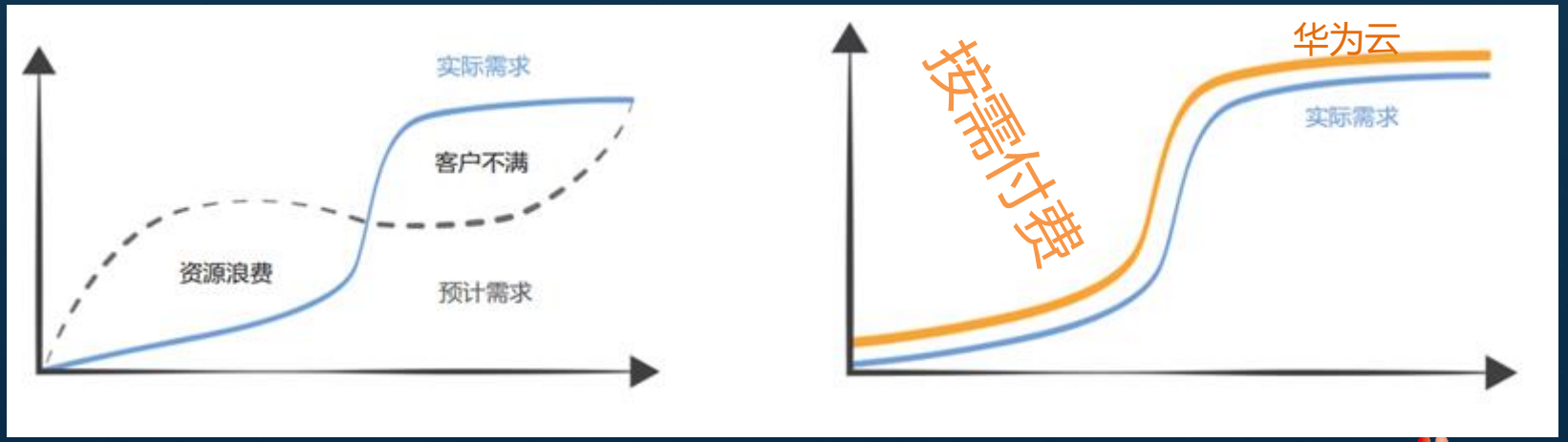
云化大数据分析平台 vs 自建大数据平台

自建大数据平台

华为云

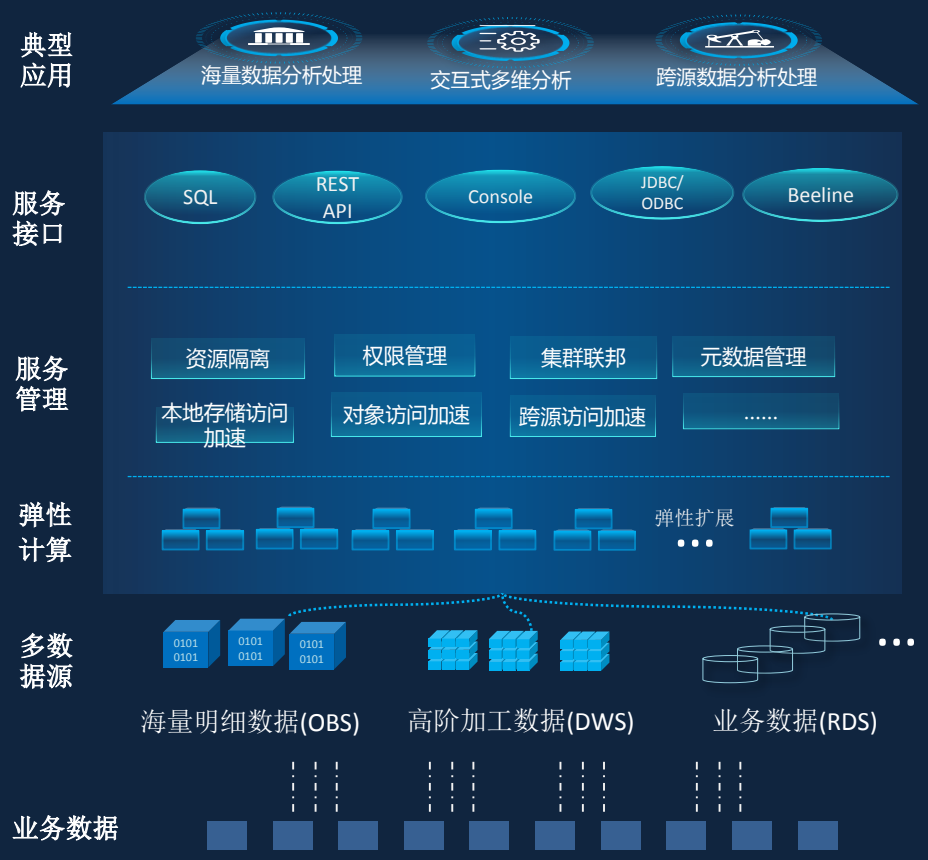
降低成本

购买机器
机房宽带
运维成本
电力成本
存储成本
计算成本



HUAWEI

DLI，完全托管的云上数据处理分析服务



两大功能

- 运行SQL作业
- 运行Spark作业

两大特点

- **异构多数据源联合分析**：支持文本/CSV/Parquet/ORC/JSON等多种数据格式，支持OBS/CloudTable/DWS/RDS跨数据源联合查询
- **SQL on AI**：使用DLI，用户可以先通过SQL直接调用AI能力（OCR识别、图像识别、图像搜索等），然后再通过SQL对返回的结果进行统计分析。

三大行业解决方案

- 基因
- 游戏
- 地理

目录

- 大数据及其应用
- 大数据技术及其流行框架
- 华为云数据湖探索（DLI）服务
- **使用DLI轻松探索数据背后的价值**

使用DLI轻松探索数据背后的价值

使用DLI轻松进行大数据ETL处理

使用DLI轻松探索数据背后的价值

使用DLI轻松进行多数据源联合分析

Thank You.