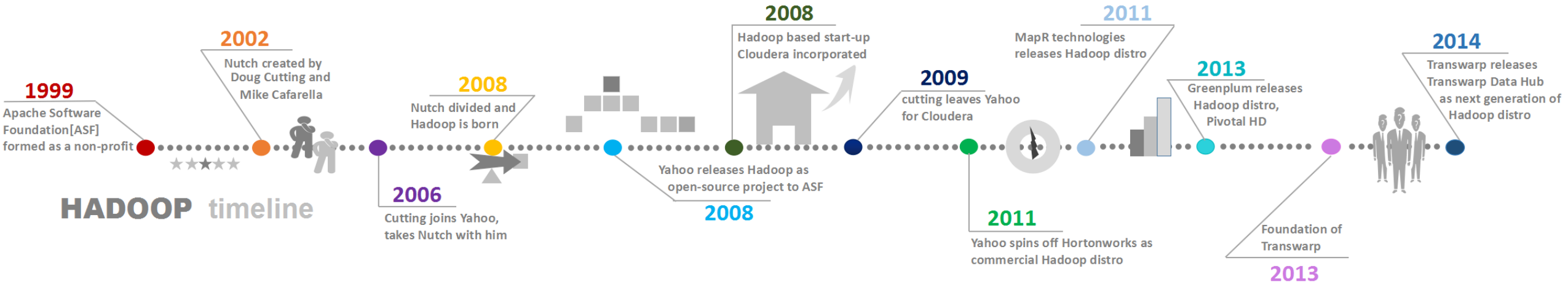
TRANSWARP
星 环 科 技

大数据技术的发展现状和最新趋势

孙元浩
星环科技
Founder & CTO
transwarp.io

Hadoop的发展历程回顾

TRANSWARP 星 环 科 技



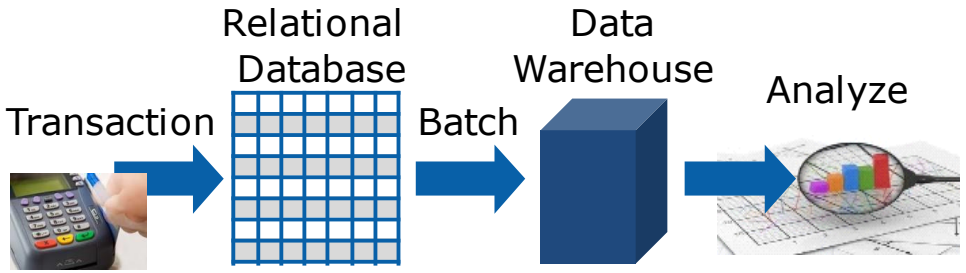
大数据技术的软件栈

TRANSWARP 星环科技

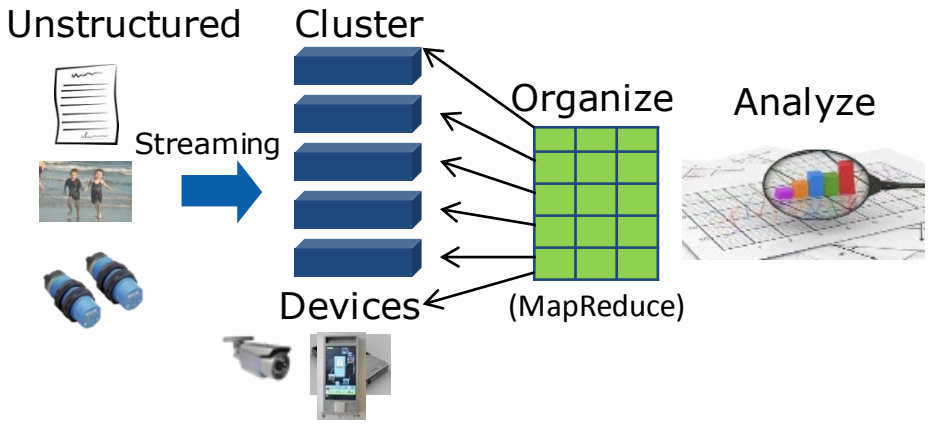
分布式计算已逐渐成为主流计算方式

TRANSWARP 星 环 科 技

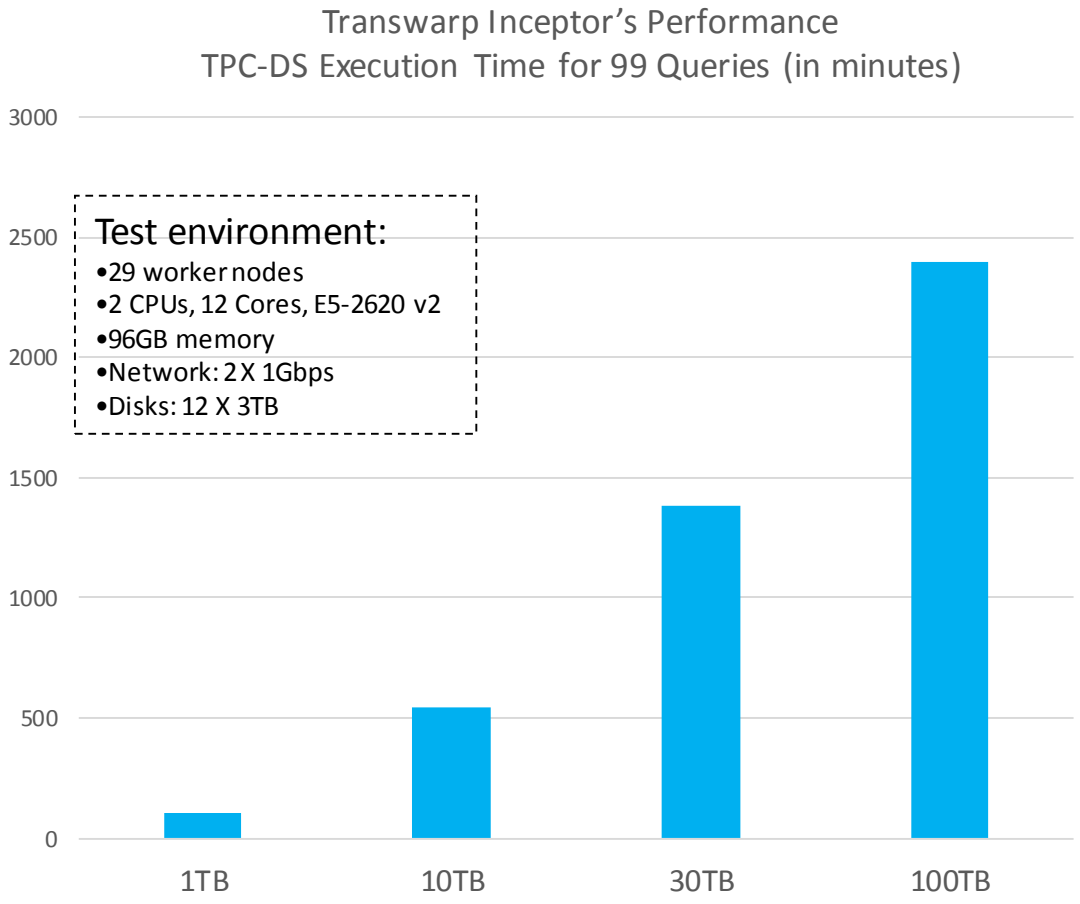
Traditional Data Analysis



Big Data Analysis



星环SQL on Hadoop已经能够高效处理100TB数据的复杂分析

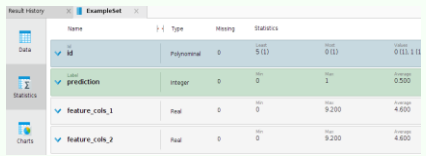
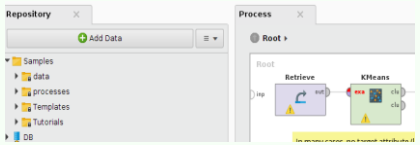
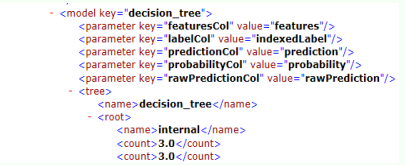


星环大数据集群已经可以在生产环境中处理20PB的数据

数据分析算法逐渐丰富，工具普及化

TRANSWARP 星环科技

- R / Python语言开发 => 算法工程师，数据科学家
- 交互式挖掘 => 业务分析师，数据科学家

数据预览	预处理	特征工程	模型训练	模型上生产
•R和Midas中可以连接TDH中数据表做数据预览 •可以对列做tag/feature的管理	•通过内置的分布式统计算法完成相关的预处理与数据分析 •支持标准化，归一化，正则化，缺失值填充，数据分箱等 •支持通过Inceptor SQL进行数据ETL处理	•结合业务领域专家知识，以及相关算法降维，选择特征指标与维度 •利用深度学习神经网络算法，通过升维降低特征工程维度选取难度	•用户通过GUI选择算法开发训练模型 •模型编译成为DAG，由Hubble组件来调度任务 •支持单机R算法和分布式算法训练模型	•导出模型为PMML •PMML模型可以转换成生产系统的代码，部署到实际业务中
		Dimension Reduction 主成分分析 Linear Regression 线性回归 Deep Learning 深度学习 ……		
rapidminer R Studio	rapidminer R Studio Inceptor		rapidminer R Studio	Discover Stream

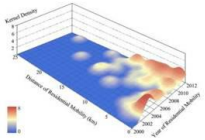
实时流处理技术推动IoT数据分析能力


ATM现场交易


飞行参数


POS刷卡


智能风机


人流信息


光伏发电


客服


炼化厂

100k records/s/node
1KB/record

分布式
队列

data mining on streams

batch @ t

batch @ t+1

SQL
aggregation

machine
learning using
java/scala/R

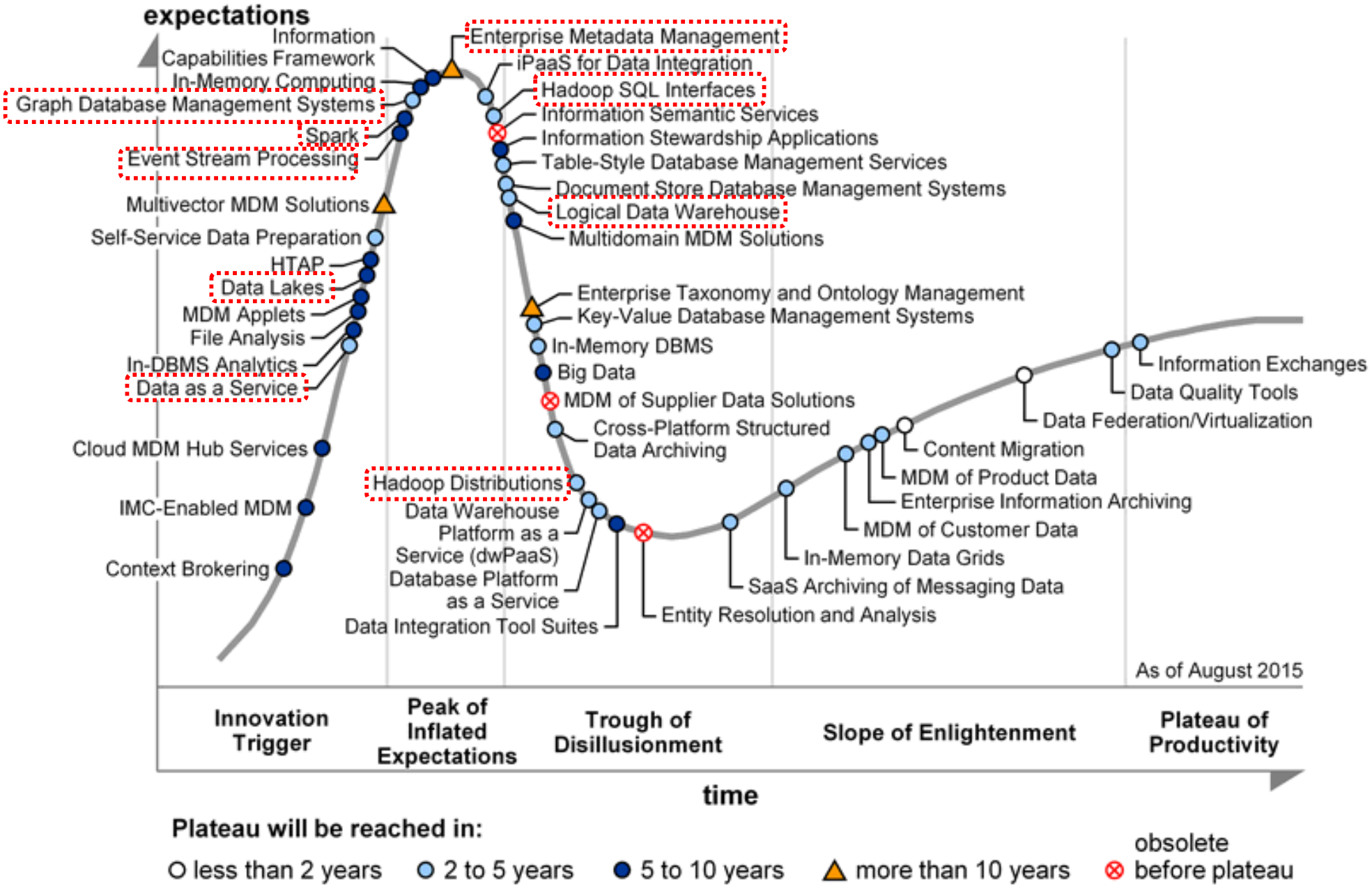
报警

transwarp stream

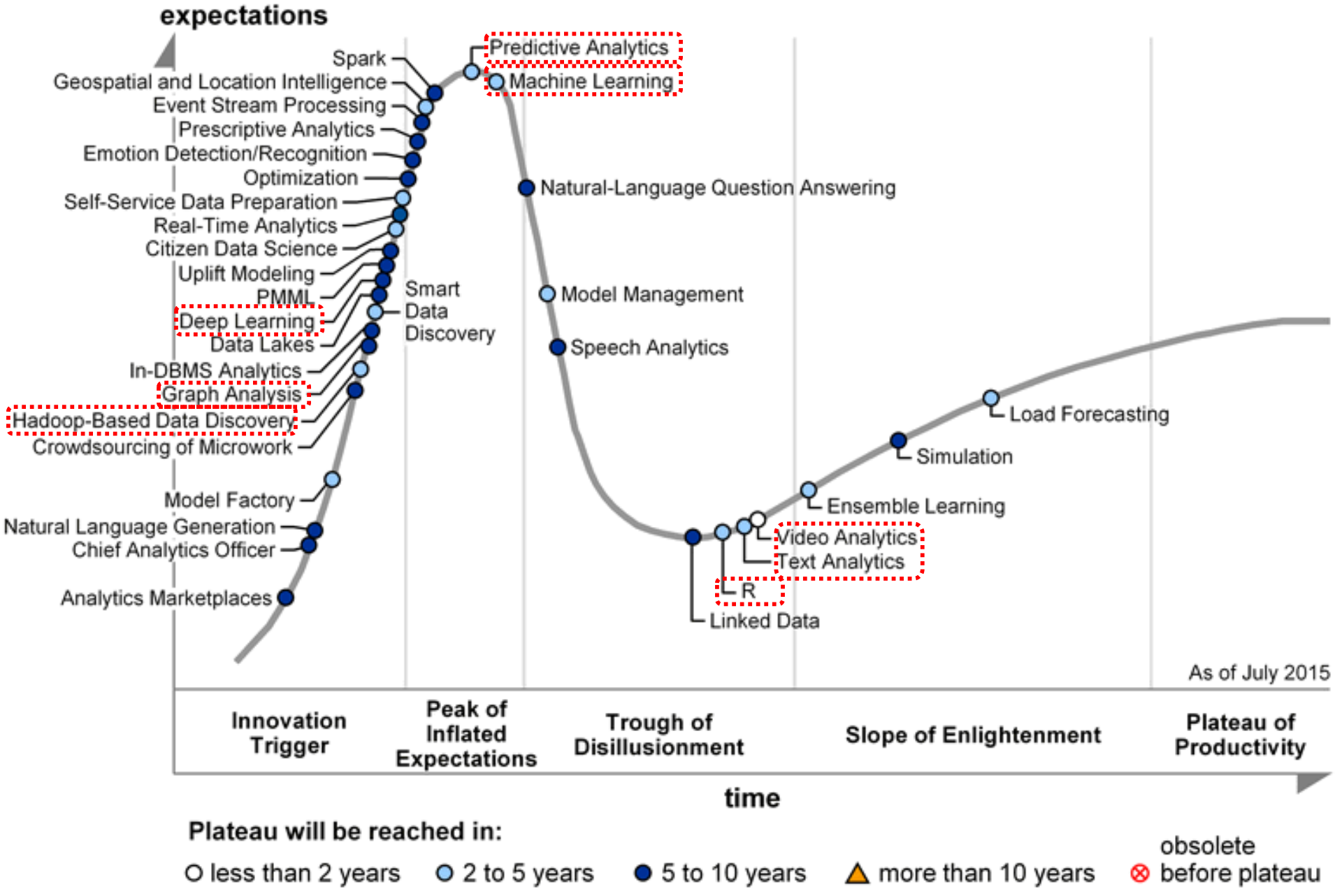
1. Streaming processing and batch processing are unified in one programming model
2. SQL and its extension is the unified declarative language for device monitoring and diagnostics.
3. ANSI SQL 2003 and PL/SQL are supported on streaming events.
4. Linear Algebra
5. Machine learning

Usage cases in IoT & FS:
Real-time event monitoring
Real-time dashboard & statistics
Real-time outlier detection
Realt-ime fraud detection

Hype Cycle for Information Infrastructure



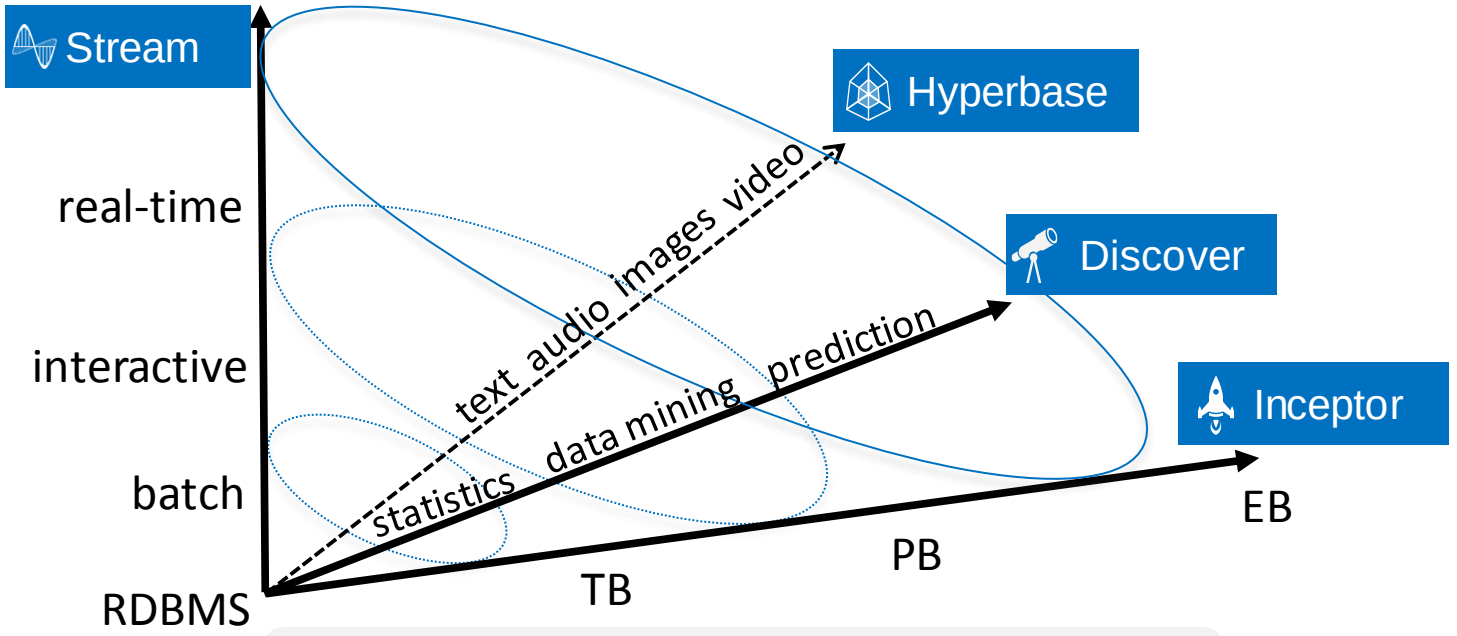
Hype Cycle for Advanced Analytics and Data Science

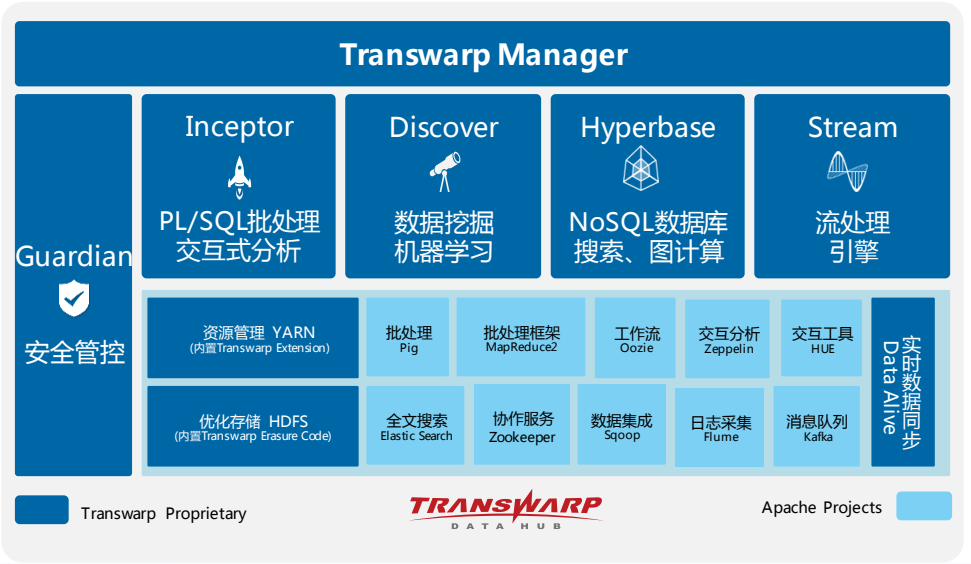


后Hadoop/Spark时代的技术发展趋势

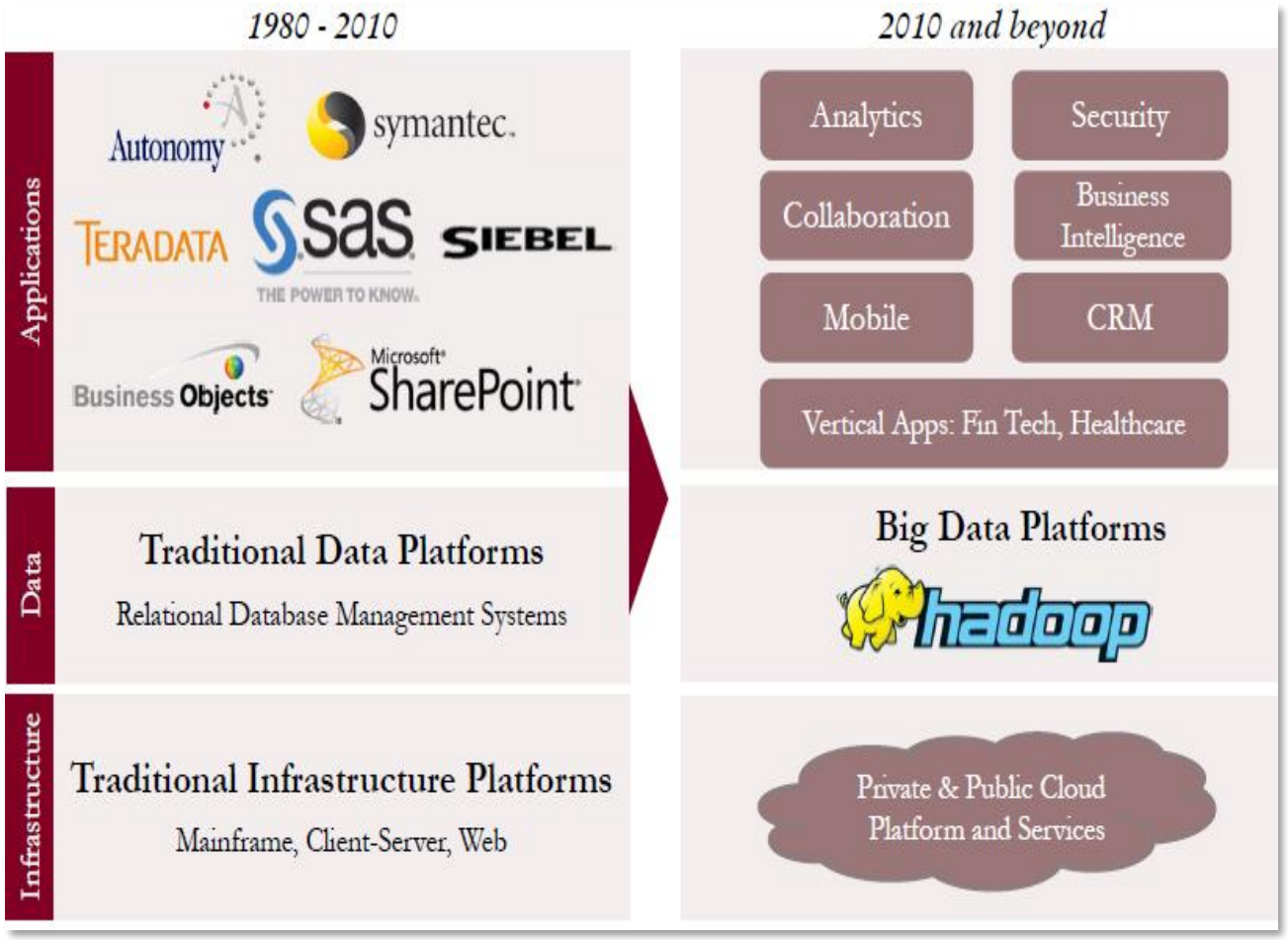
TRANSWARP 星环科技

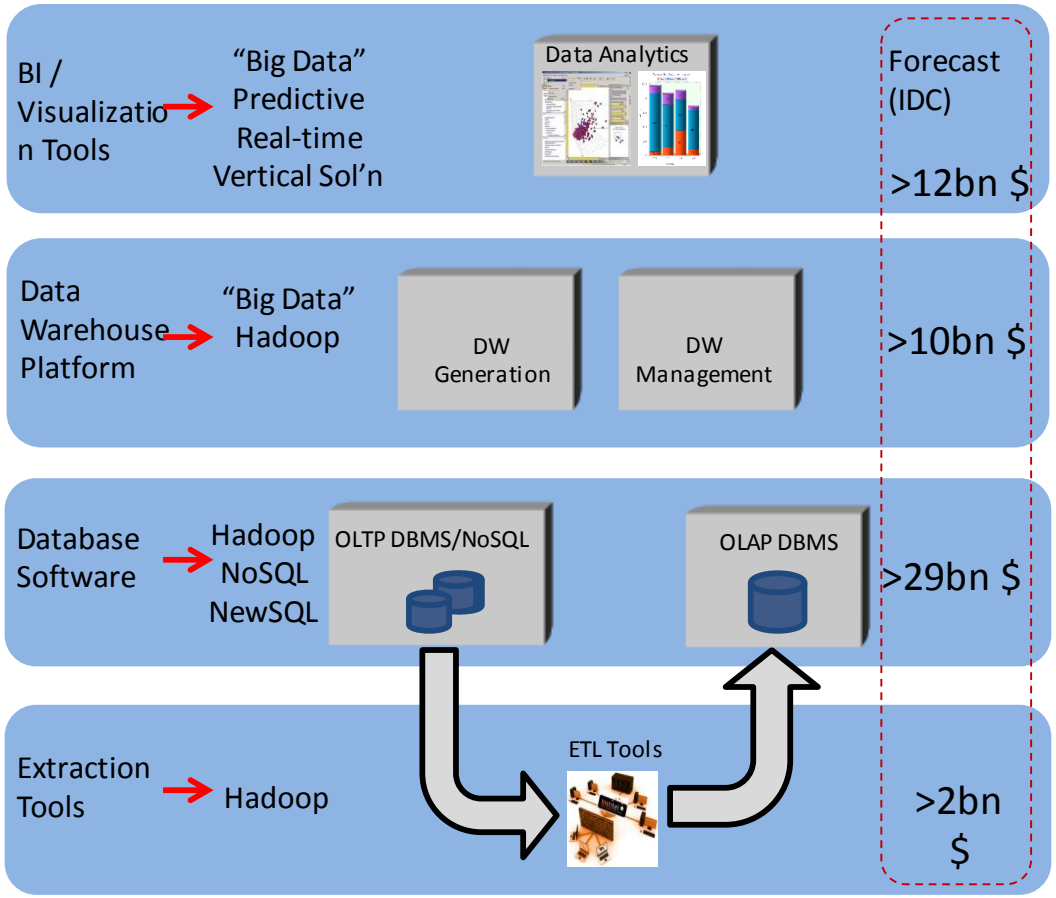
新技术在四个维度上加速创新	
数据量 Volume	驱动力来自结构化数据的不断增加，企业需要性价比更高的技术方案
	SQL on Hadoop快速成熟，Big Data Database将替代传统relational database。传统的MPP分析型数据库将消失。
类型 Variety	驱动力来自对多种数据类型数据（文本、图片、音频、视频）的存储和分析需求
	深度机器学习技术（如TensorFlow）快速发展并得到应用，基于GPU／FPGA的加速技术逐渐普及。
速度 Velocity	从离线处理进步到实时数据处理，特别是IoT的广泛部署，推动了对实时计算的需求
	Flink，Apex，SqlStream，Internana，ParStream，Transwarp Stream等新技术，融合了批处理和流处理，提供强大易用的低延时实时计算能力，将逐渐取代现有流处理技术。
价值 Value	从历史统计发展到预测性分析。大数据的真正意义在于从数据中发现价值。
	数据挖掘、机器学习、图计算等产品和工具将日益普及，使用门槛将极大降低，普通业务人员很快能够自助进行分析建模。





Hadoop及其生态系统将重构数据处理市场

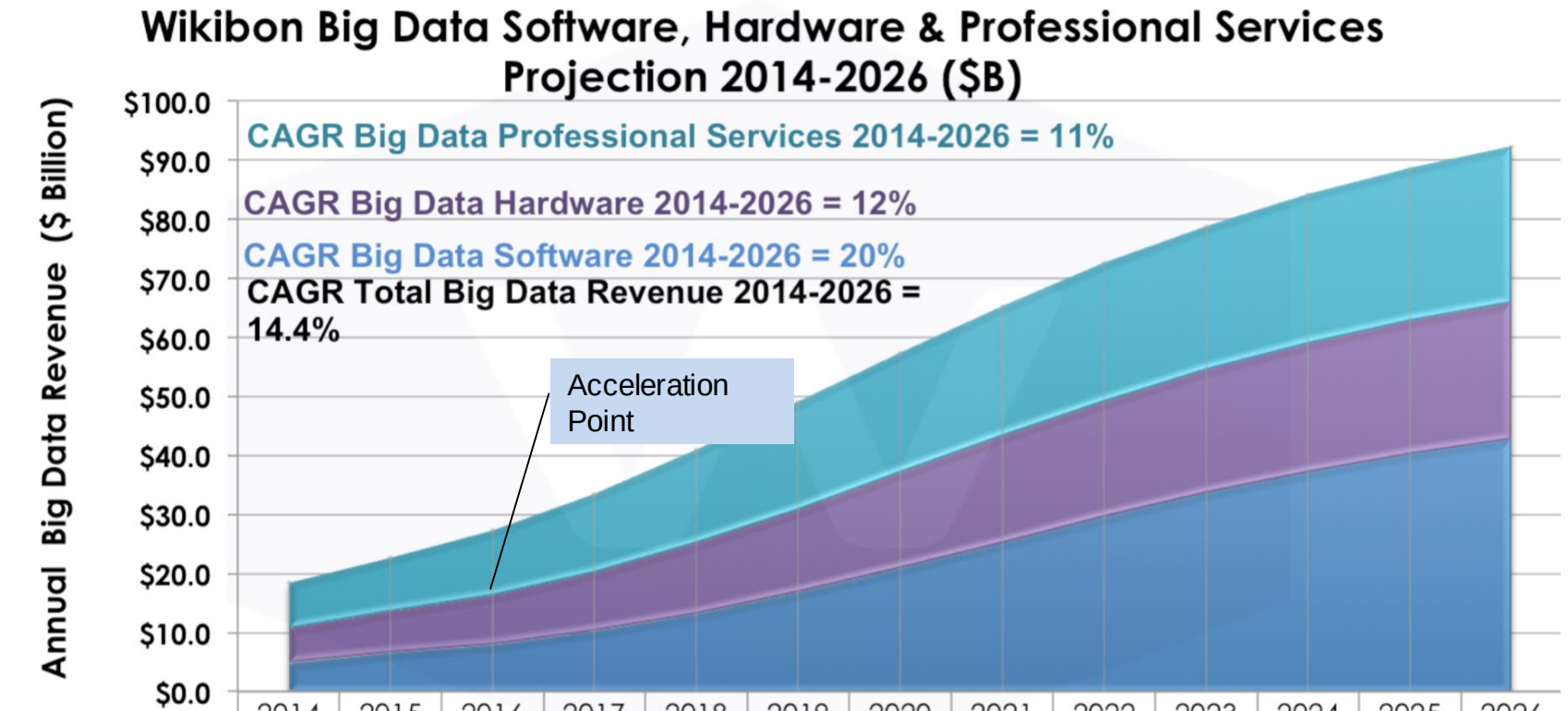




大数据产业链日益繁荣


专业服务、定制开发
大数据应用软件
分析软件、模型
可视化、工具等
大数据
平台
软件
OPERA SOLUTIONS
IBM
accenture 埃森哲
Deloitte. 德勤
宇信科技 Yusys Technologies
Palantir
guavus
EVERSTRING
bloomreach
OPOWER
zest finance
LendingClub
AYASDI
ThreatMetrix
23andMe
tableau
Qlik
Microsoft Power BI
platfora
GoodData
TRIFACTA
R Studio
rapidminer
TRANSWARP DATA HUB
sas
SPSS
H2O.ai
CONTINUUM ANALYTICS
TRANSWARP DATA HUB
cloudera
MAPR TECHNOLOGIES
Hortonworks
databricks
TERADATA
Pivotal
IBM
SAP HANA
Qubole
bluedata

2016年是Hadoop技术大规模应用的战略转折点



	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
Professional Services	$7.6	$9.1	$11.1	$13.4	$15.8	$18.2	$20.3	$22.0	$23.3	$24.3	$25.1	$25.8	$26.3
Big Data Hardware	$5.7	$6.9	$8.2	$9.9	$11.8	$14.0	$16.1	$18.0	$19.6	$20.8	$21.8	$22.5	$23.1
Big Data Software	$4.9	$6.6	$8.0	$10.2	$13.2	$16.8	$20.9	$25.2	$29.5	$33.5	$37.1	$40.2	$42.7
Total Big Data	$18.3	$22.6	$27.3	$33.5	$40.8	$49.0	$57.3	$65.2	$72.4	$78.7	$84.0	$88.5	$92.2

Source: © Wikibon Big Data Project, 2016

未来五年大数据市场将以每年30%的速度增长


TRANSWARP