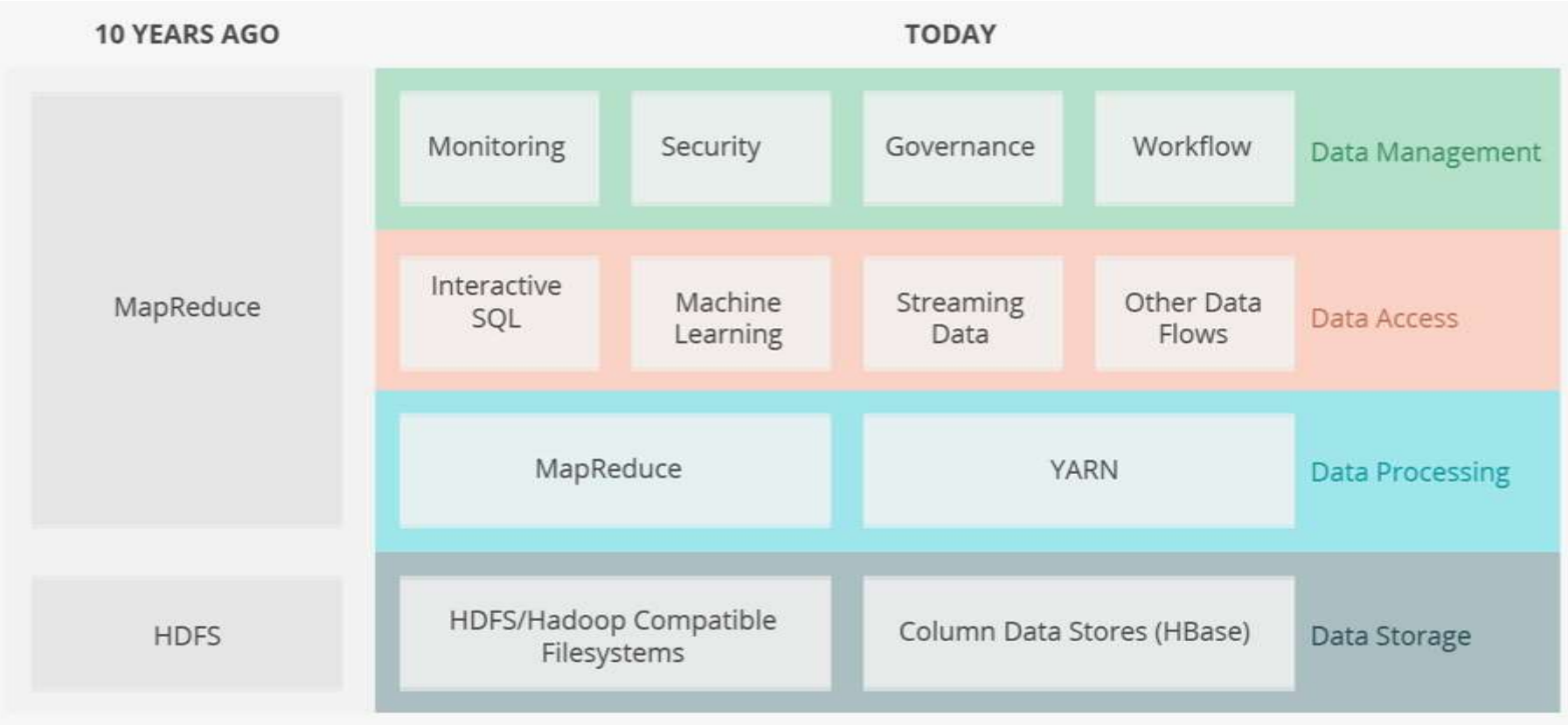
Зачем нужен «стандартный» HADOOP?

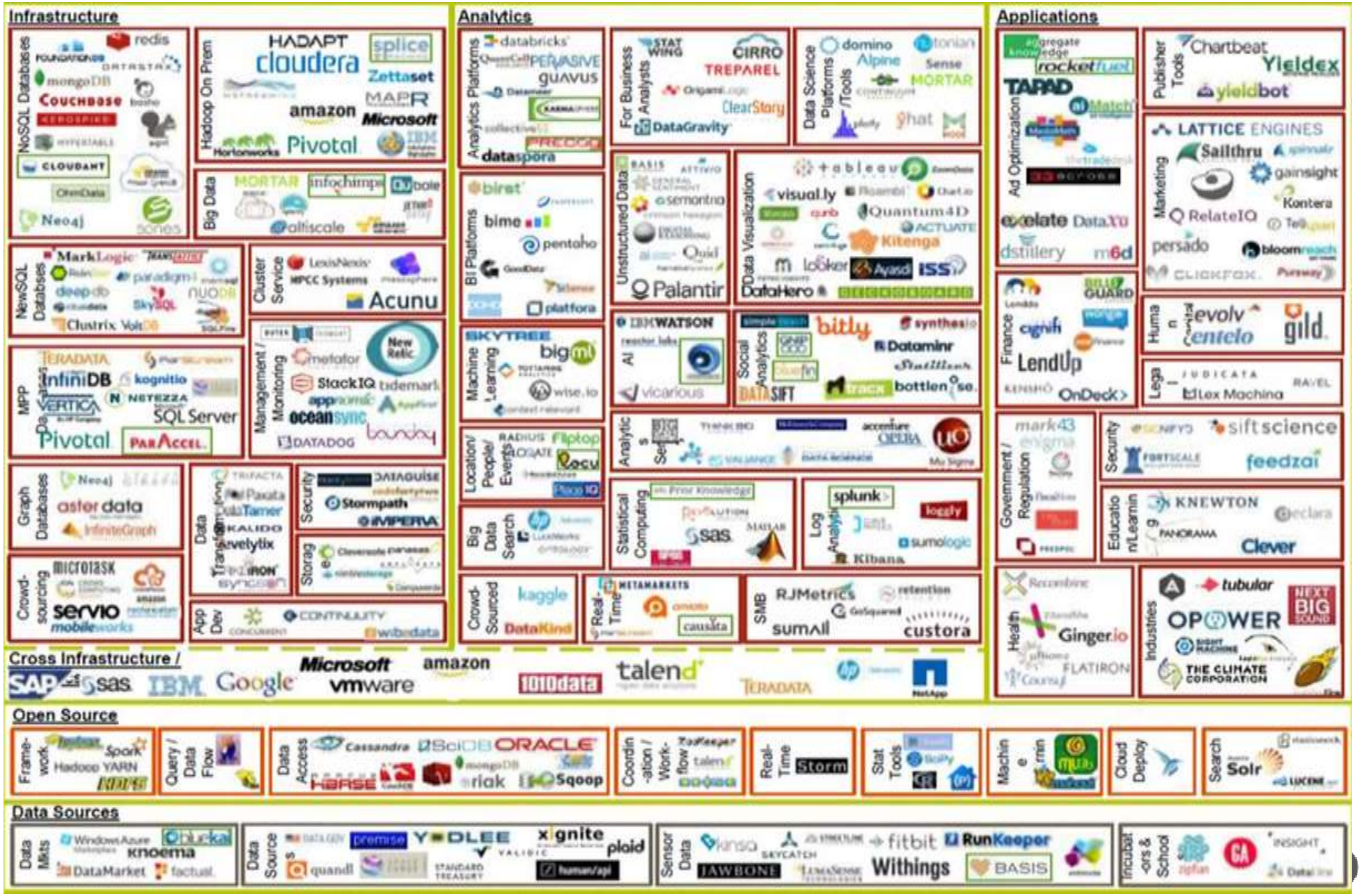
Сергей Золотарев
Arenadata

ARENADATA

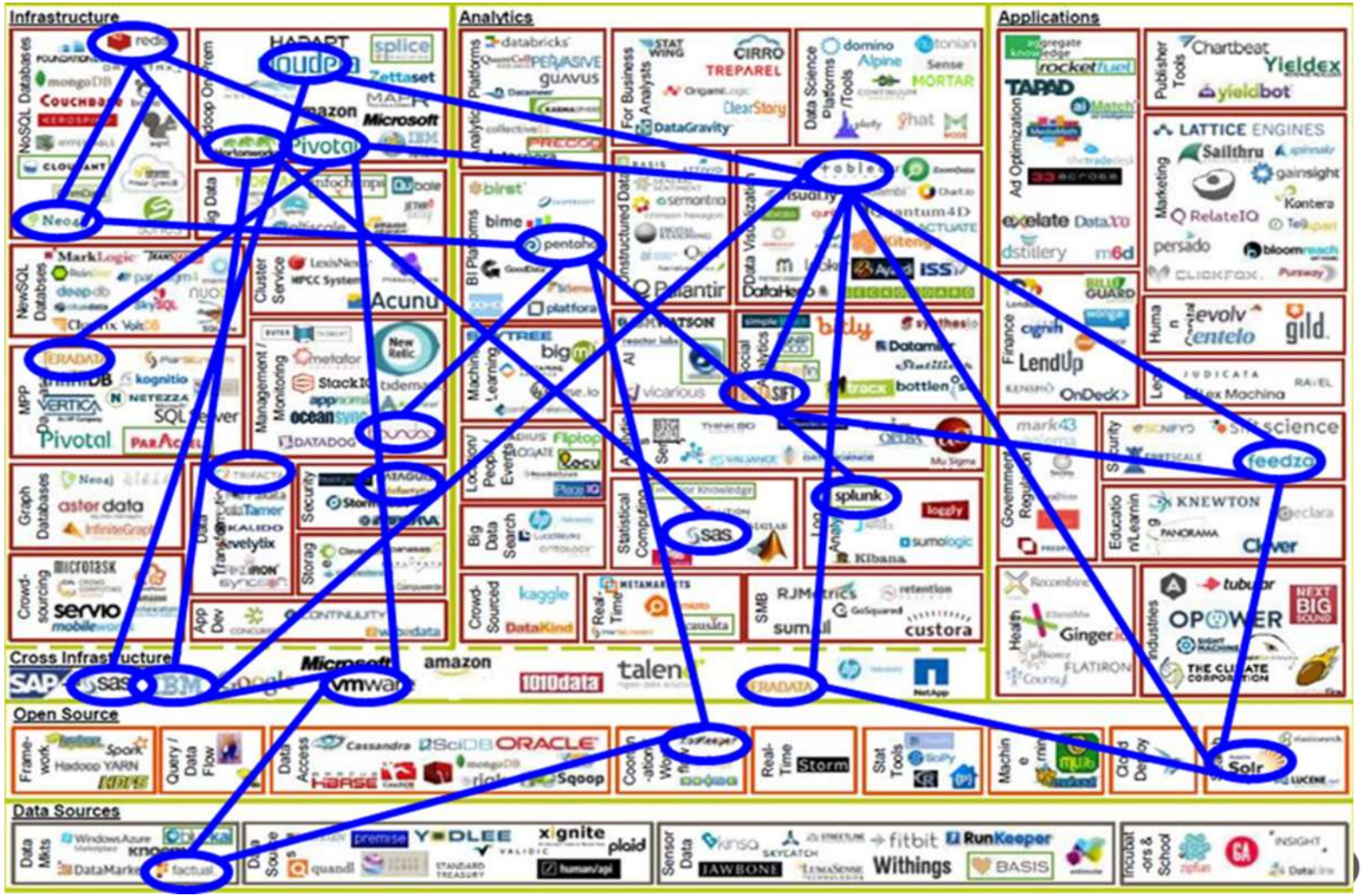
Развитие экосистемы Hadoop



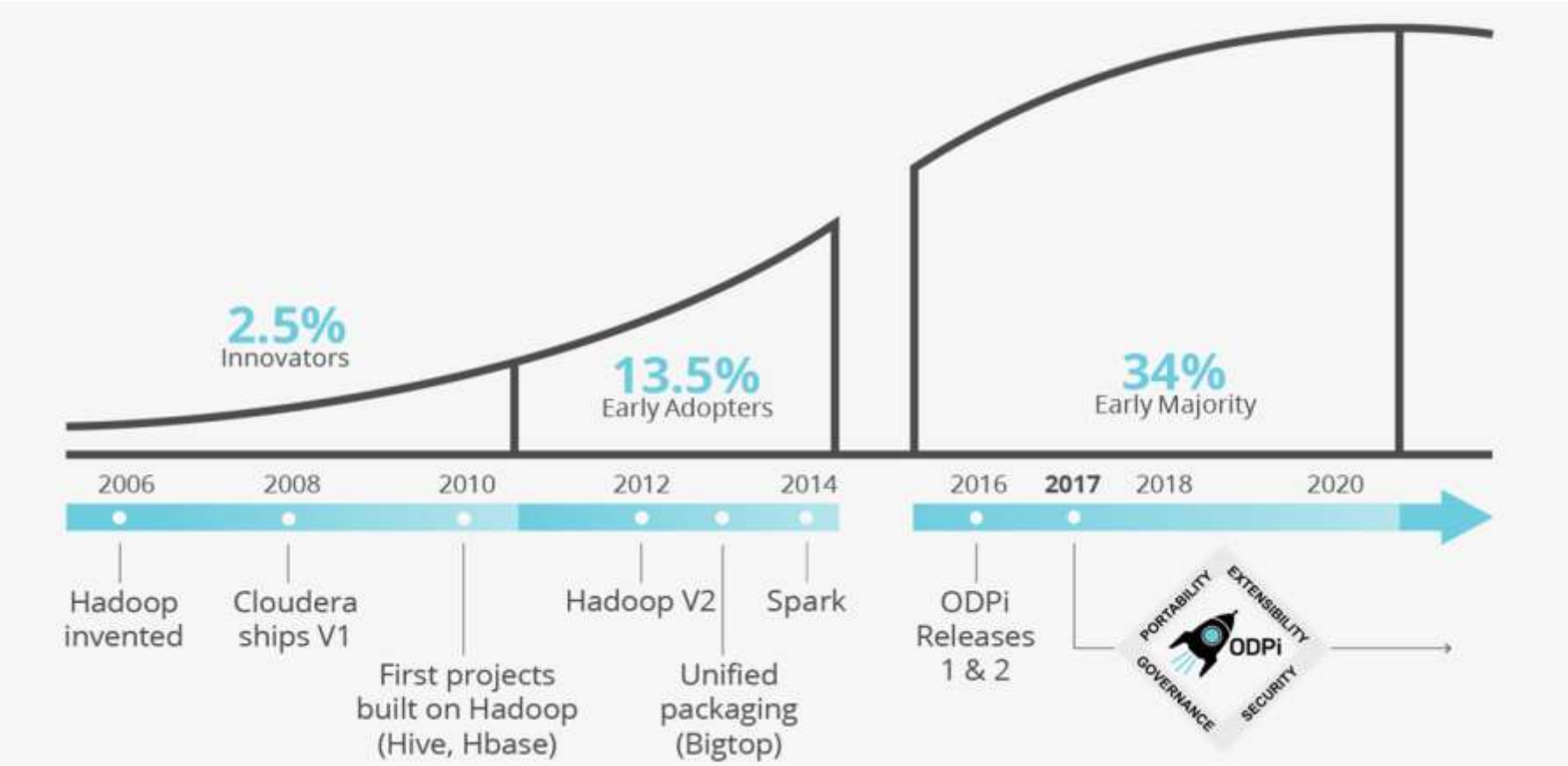
ARENADATA


Infrastructure
Analytics
Applications
NoSQL Databases
Hadoop On Prem
Big Data
NewSQL Databases
Cluster Service
MPP Databases
Management / Monitoring
Graph Databases
Data Transformation
Security
Crowd-sourcing
Storage
App Dev
Analytics Platforms
For Business Analysts
Data Science Platforms /Tools
BI Platforms
Unstructured Data
Data Visualization
Machine Learning
AI
Social Analytics
Location/ People/ Events
Analytic Services
Big Data Search
Statistical Computing
Log Analytics
Crowd-Sourced
Real-Time
SMB
Ad Optimization
Publisher Tools
Marketing
Finance
Legal
Government / Regulation
Security
Education/ Learning
Health
Industries
Cross Infrastructure / Analytics
Open Source
Framework
Query / Data Flow
Data Access
Coordination / Workflow
Real-Time
Stat Tools
Machine Learning
Cloud Deploy
Search
Data Sources
Data Mkts
Data Sources
Sensor Data
Incubators & Schools

ATA

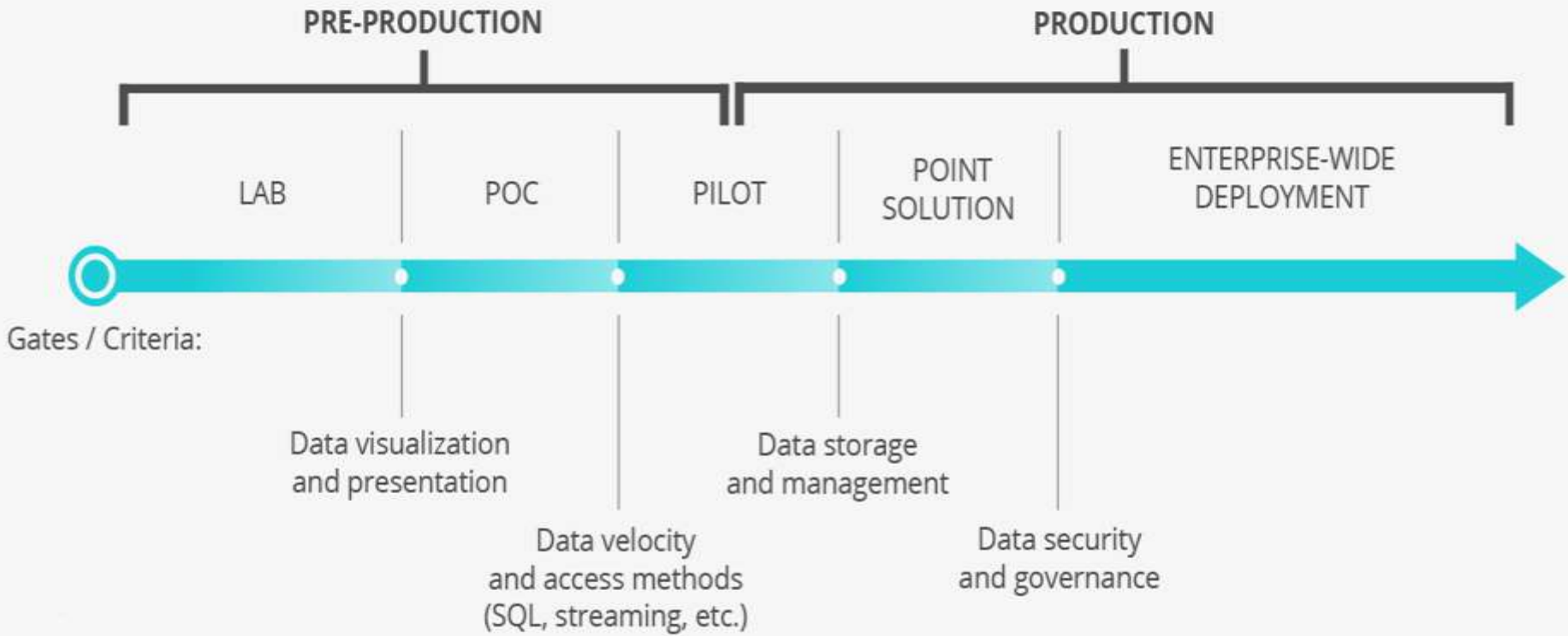

Infrastructure
Analytics
Applications
Cross Infrastructure
Open Source
Data Sources
NoSQL Databases
NewSQL Databases
MPP
Graph Databases
Crowd-sourcing
Cluster Service
Management / Monitoring
Security
Storage
App Dev
Platforms
For Business Analysts
Data Science Platforms /Tools
BI Platforms
Machine Learning
Location/ People/ Events
Big Data Search
Statistical Computing
Crowd-Sourced
Real-Time
SMB
Ad Optimization
Marketing
Finance
Government Regulation
Security
Education/Learning
Health
Industries
Publisher Tools
Human
Framework
Query / Data Flow
Data Access
Coordination / Workflow
Real-Time
Stat Tools
Machine
Cloud Deploy
Data Marts
Data Sources
Sensor Data
Incubators & School

Как корпорации «привыкают» к Hadoop



ARENADATA

Только 28% проектов в «Продакшн»



ARENADATA

Что нужно корпоративному ИТ

SECURITY	DATA GOVERNANCE	OPERATIONS
Authenticate	Profile	Provision
Authorize	Classify	Configure
Audit	Collaborate	Manage/Upgrade
Setup Policies & Entitlements	Understand Quality	Monitor
Protect	Leverage Metadata	Scale
Understand Risk Profile	Provenance & Lineage	Perform

ARENADATA

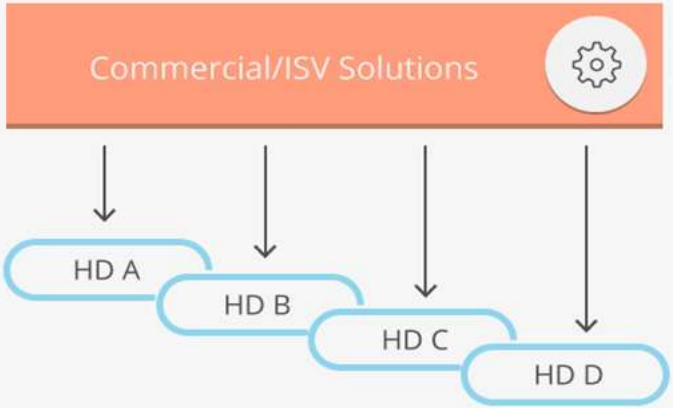
Принципиальное отличие Пилота от Продакшн

HADOOP/BIG DATA OPERATIONAL CONSIDERATIONS	LAB AND LIMITED PRODUCTION (PILOT, POC)	ENTERPRISE-WIDE PRODUCTION
Deployment type (single cloud, multi-cloud, on prem, single vendor, multi-vendor, etc.)	Whatever is cheapest and easiest	Tied to corporate and/or departmental procurement guidelines, and must consider SLAs, regulations, staffing, etc.
Security and data governance (data access, protection, storage, transmission, etc.)	PoCs often limit data set scope in order to keep things simple to get stakeholder buy in.	Biggest strategic gains will involve data that must adhere to strict security protections, such as customer preferences, and regulatory requirements that often require policy portability and that vary internationally.
Software Lifecycle Management (updating, monitoring and managing applications and platforms)	Limited usage scope means manageability and governance issues can be contained	As the number and type of uses, and the number and skill level of users, expand, the need for consistent approaches to version control, updates, application deployment, data access and other governance issues increase.
Big Data / Analytics Applications (ensuring uptime, extensibility and portability)	Most PoC stacks are purpose built to support a single application, so minimal need for portability and extensibility; single stack also minimizes dependency issues	With varied use-cases and tools working across shared data lakes, enterprises require a set of disciplined practices to support the objectives of each stakeholder.

Стандартизация снижает сложность внедрения

WITHOUT ODPI

Multi-distro certifications and regression testing increases ISV development, burden, and enterprise support costs



Multiple Hadoop Distribution Compliance Required

Duplicate effort

*HD = Hadoop Distribution

WITH ODPI

ODPi Interoperable Solutions

sas Pivotal IBM XAVIENT
DATATORRENT syncsort wanDISCO

ODPi RUNTIME CERTIFIED

ODPi INTEROPERABLE

ODPi Runtime Compliant Platforms

ARENADATA AsiaInfo 亚信数据 altiscale
HORTONWORKS Infosys IBM

ODPi Runtime Specification

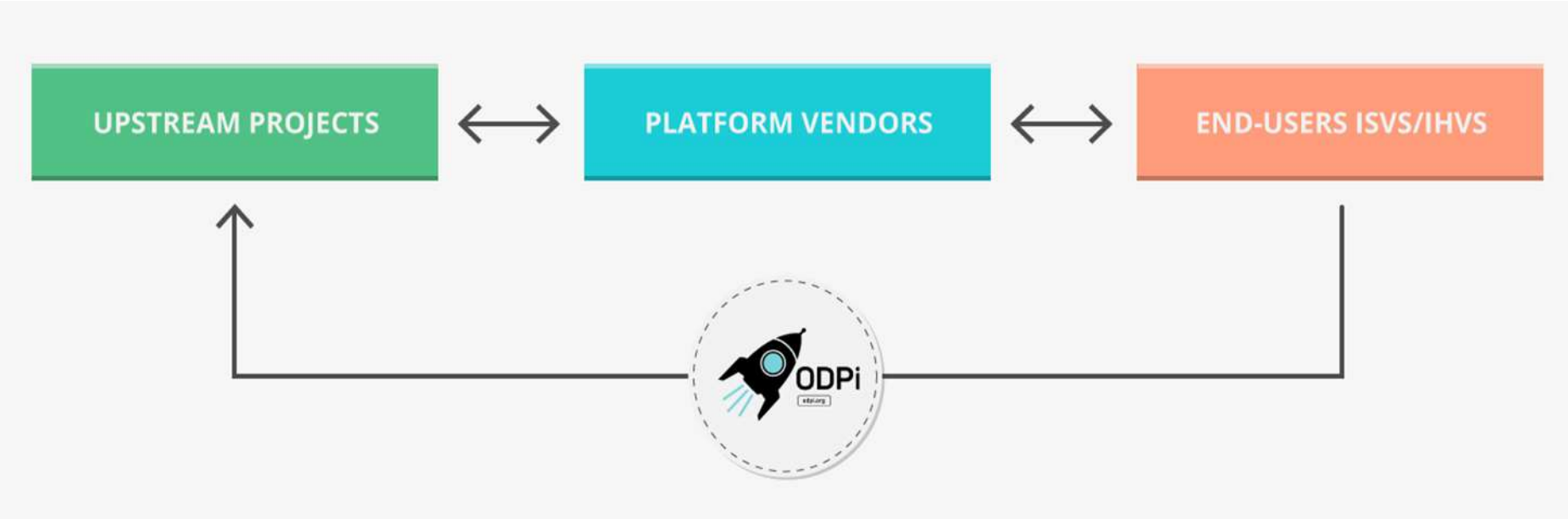
HDFS YARN MapReduce HCFS Hive

ARENADATA

ODPi – крупнейшее мировое сообщество в области больших данных

ARENADATA

Вовлеченность – залог успеха



Arenadata Hadoop : Основные факты

- ARENADATA HADOOP (ADH) – корпоративный дистрибутив платформы хранения данных Apache Hadoop
- Построен полностью на открытых компонентах проекта Apache
- Российское программное обеспечение
- Сертифицирован в 2016 году ODPi Runtime Compliant

ARENADATA

All the following Apache Hadoop platforms are ODPi Runtime Compliant. This dramatically decreases engineering complexity for Big Data developers by ensuring a consistent set of base level expectations.

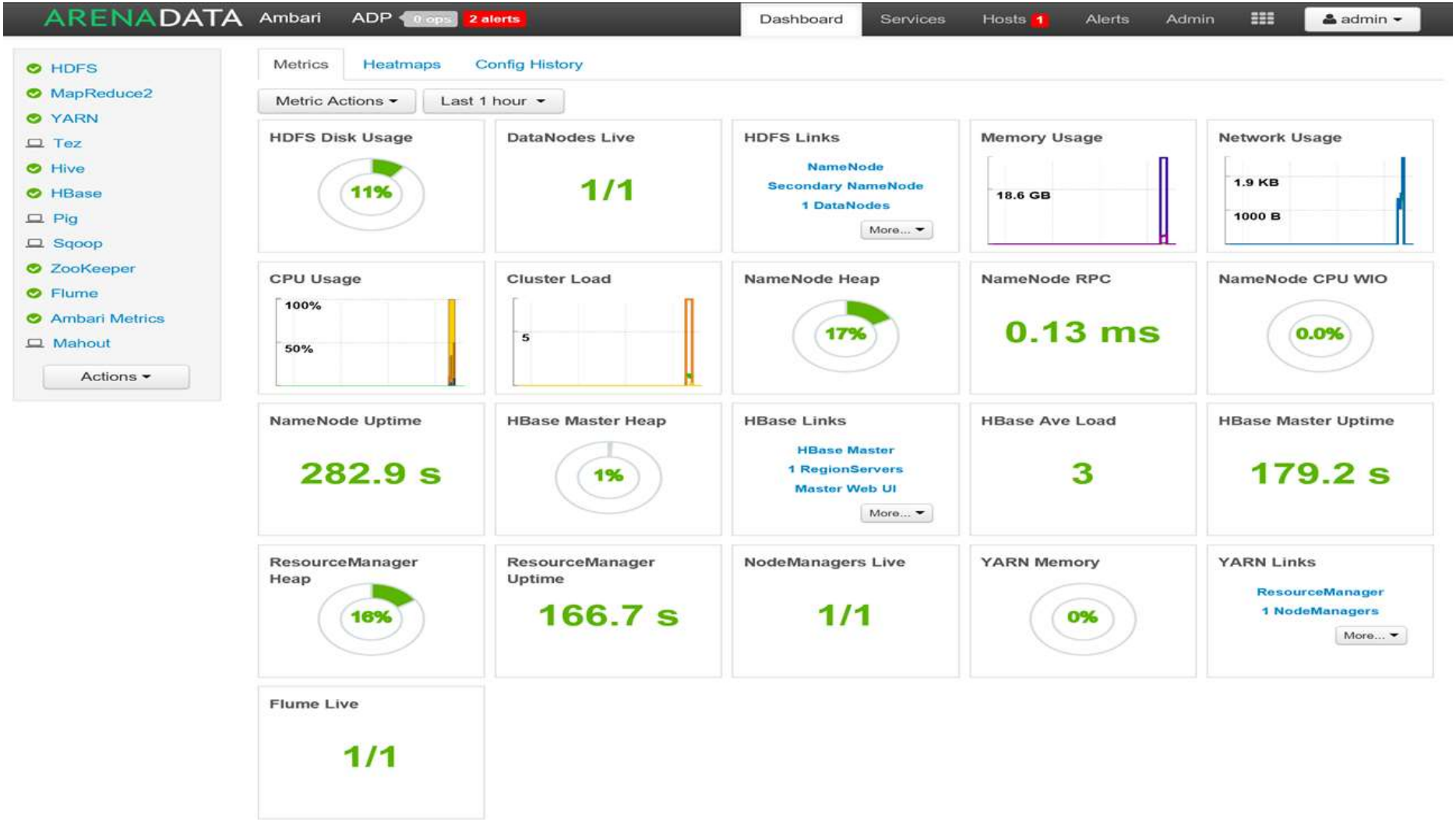
VENDOR	PRODUCT / VERSION	CONTACT
altiscale	Altiscale Data Cloud 4.2	Raymie Strata <rstata@altiscale.com>
ARENADATA	Arenadata Hadoop (ADH) 1.3.1	Alexander Nermakov <ean@arenadata.io>
HORTONWORKS	HDP 2.4.2	Alan Gates <gates@hortonworks.com>
IBM	IOP 4.2 / BigInsights 4.2	Susan Malaika <malaika@us.ibm.com>

Основные особенности

- Вся поддержка и экспертиза доступна в России и на русском языке, как удаленно так и on-site
- Разработан пакет утилит для оффлайн установки (без доступа в интернет);
- Эксплуатационная документация на русском языке

- Доступен не только в виде ПО , но и как Hadoop Appliance с полной и единой поддержкой всего программно-аппаратного комплекса от вендора
- Есть набор доступных типовых пакетных сервисов по планированию, установке и аудиту системы
- Поддержка ISV на этапе разработке решения/приложения

ARENADATA


ARENADATA
Ambari
ADP
0 ops
2 alerts
Dashboard
Services
Hosts 1
Alerts
Admin
admin
HDFS
MapReduce2
YARN
Tez
Hive
HBase
Pig
Sqoop
ZooKeeper
Flume
Ambari Metrics
Mahout
Actions
Metrics
Heatmaps
Config History
Metric Actions
Last 1 hour
HDFS Disk Usage
11%
DataNodes Live
1/1
HDFS Links
NameNode
Secondary NameNode
1 DataNodes
More...
Memory Usage
18.6 GB
Network Usage
1.9 KB
1000 B
CPU Usage
100%
50%
Cluster Load
5
NameNode Heap
17%
NameNode RPC
0.13 ms
NameNode CPU WIO
0.0%
NameNode Uptime
282.9 s
HBase Master Heap
1%
HBase Links
HBase Master
1 RegionServers
Master Web UI
More...
HBase Ave Load
3
HBase Master Uptime
179.2 s
ResourceManager Heap
18%
ResourceManager Uptime
166.7 s
NodeManagers Live
1/1
YARN Memory
0%
YARN Links
ResourceManager
1 NodeManagers
More...
Flume Live
1/1

WWW.ARENADATA.IO

ARENADATA