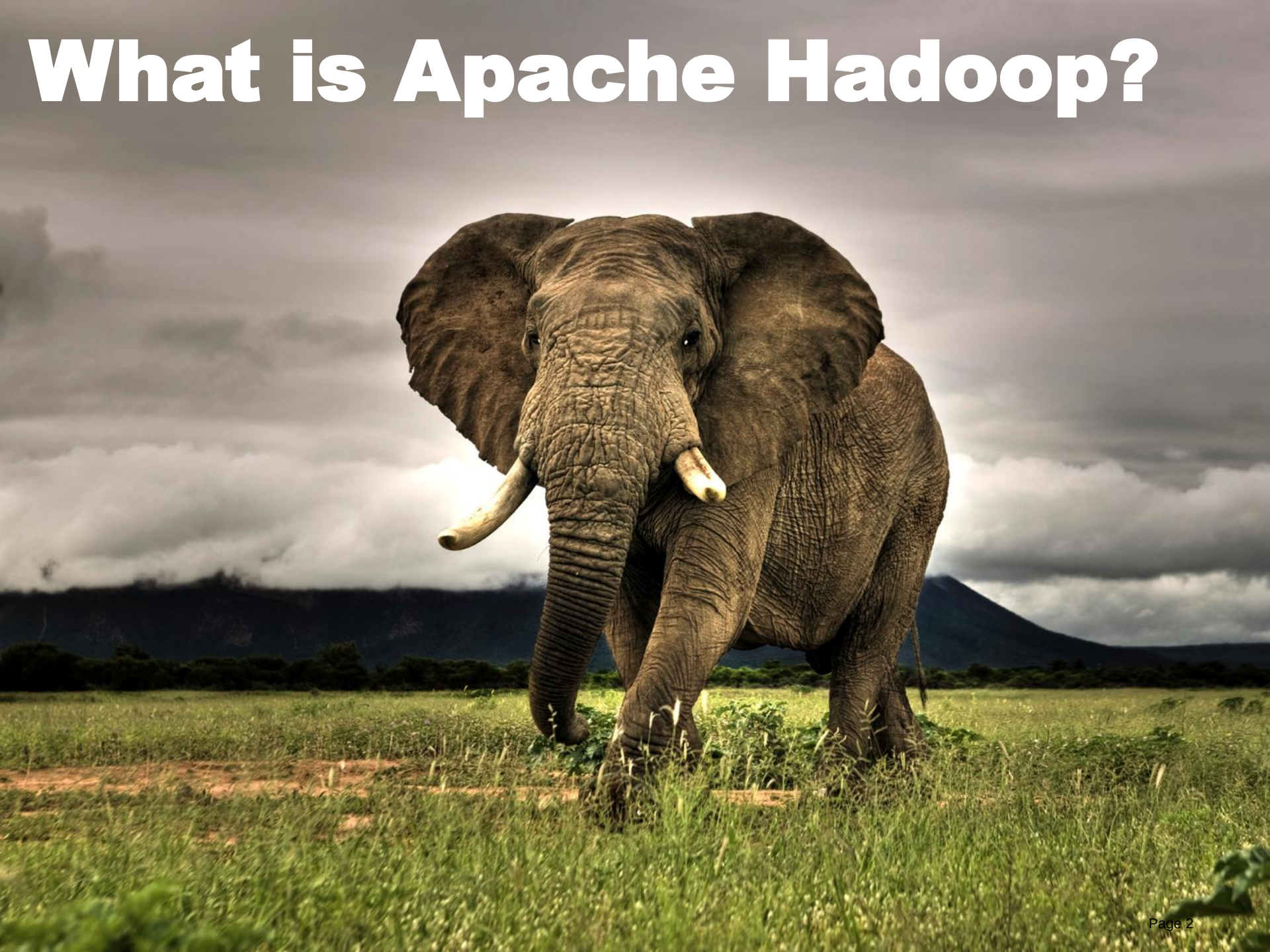
Hortonworks

Apache Hadoop Present & Future

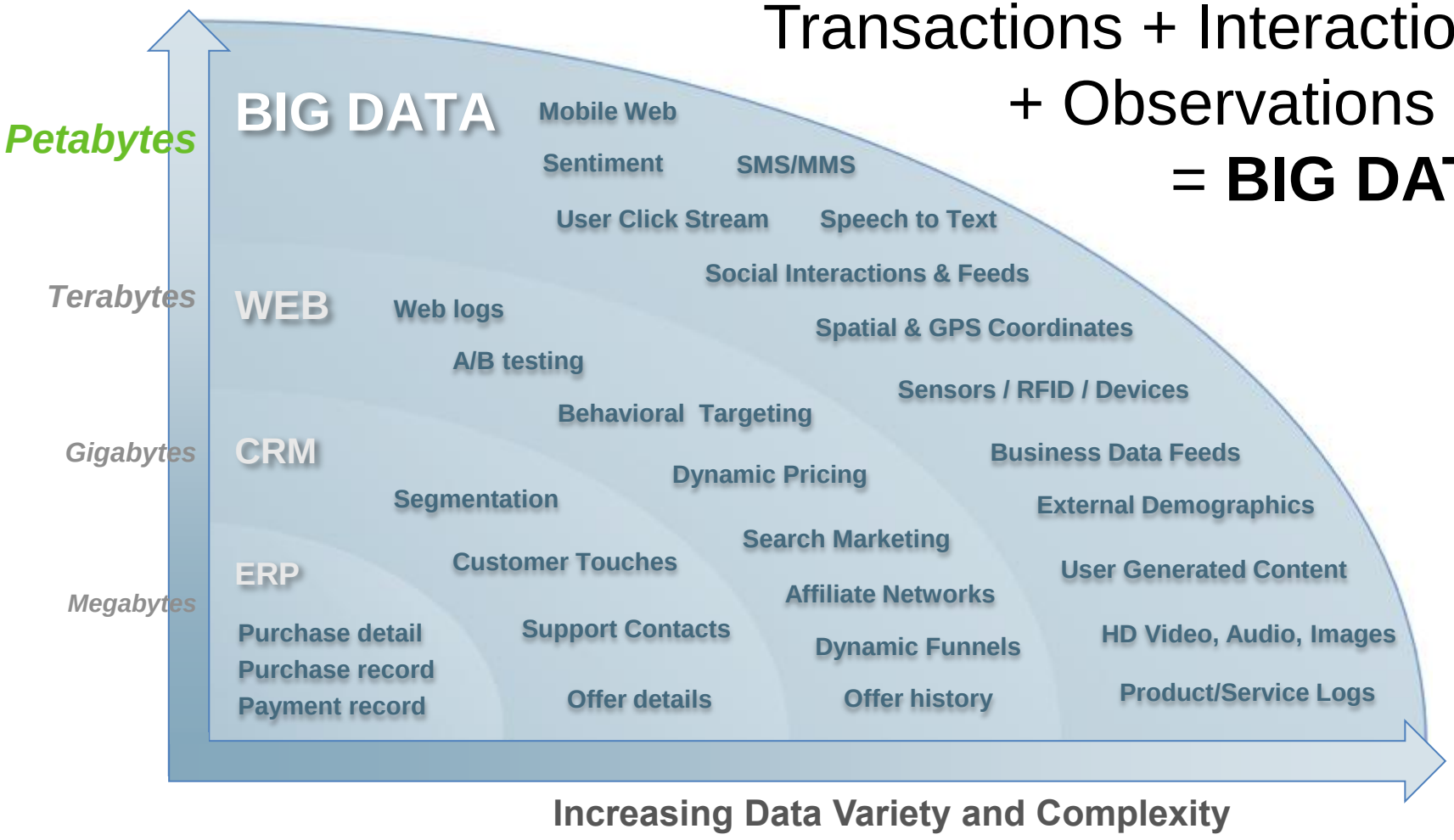
Hadoop in China 2012

Eric Baldeschwieler
CTO Hortonworks
Twitter: @eric14 , @Hortonworks

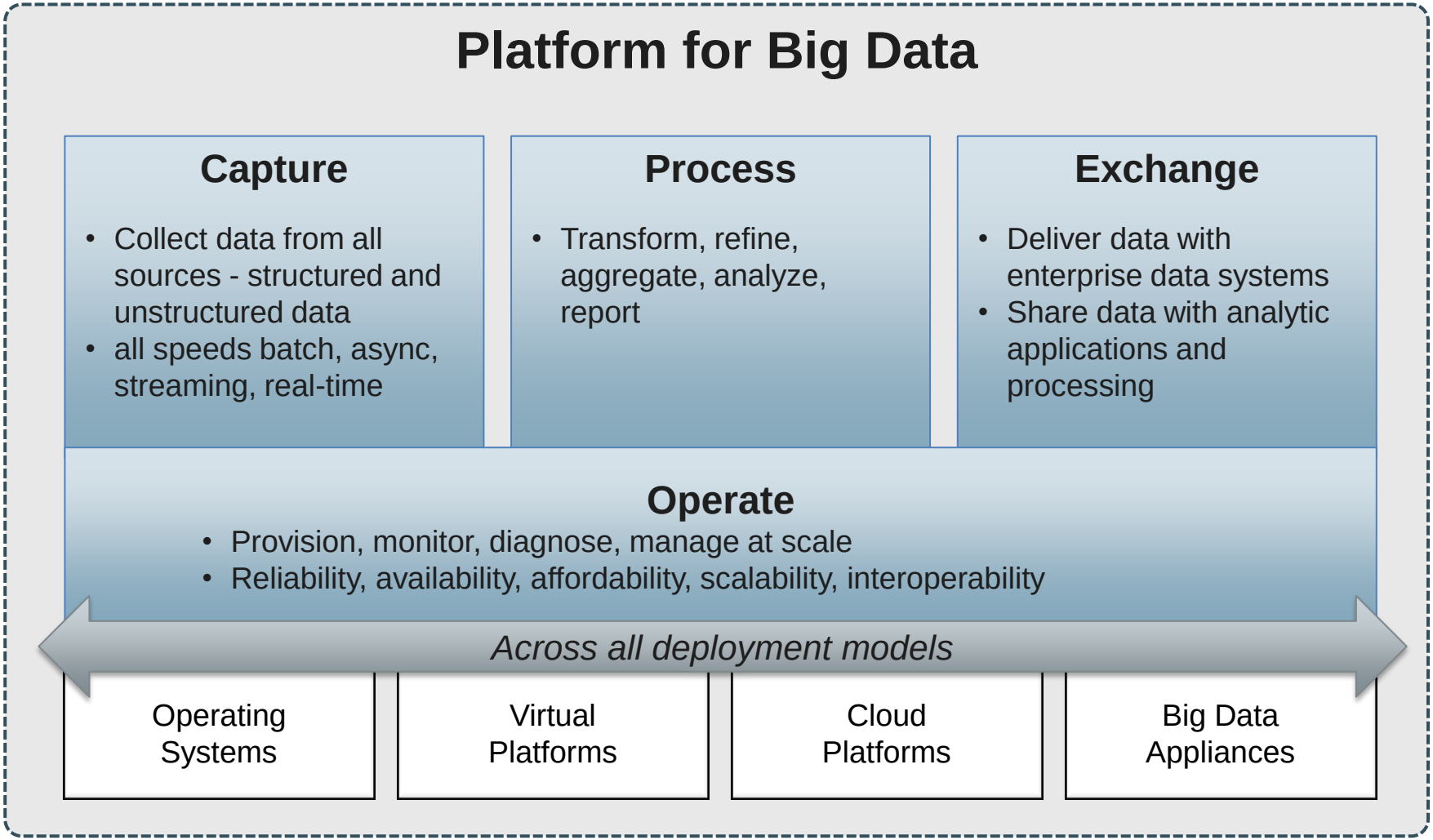
What is Apache Hadoop?



Big Data



The Big Data Problem



Big Data Use Cases

Vertical	Refine	Explore	Enrich
Social and Web	• MDM • CRM • Ad service models	• Paths • Feature usefulness	• Friends and associations • Content recommendations • Ad service
Retail	• Loyalty programs • Cross-channel customer • MDM • CRM	• Referrers • Brand and Sentiment Analysis • Paths • Taxonomic relationships	• Dynamic Pricing/Targeted Offer
Intelligence	• Threat Identification	• Person of Interest Discovery	• Cross Jurisdiction Queries
Finance	• Risk Modeling & Fraud Identification • Trade Performance Analytics	• Surveillance and Fraud Detection • Customer Risk Analysis	• Real-time upsell, cross sales marketing offers
Energy & Utility	• Production Optimization • Smart Meters & Devices	• System wide analysis • Fine grained reporting	• Automatic system recovery • Interactive customer service
Manufacturing	• Supply Chain Optimization	• Customer Churn Analysis	• Dynamic Delivery • Replacement parts
Healthcare & Payer	• Electronic Medical Records (EMPI)	• Clinical Trials Analysis	• Insurance Premium Determination

Apache Hadoop, Big Data Platform

Open Source data management with scale-out storage & distributed processing

Storage

HDFS

- Distributed across “nodes”
- Natively redundant
- Name node tracks locations

Processing

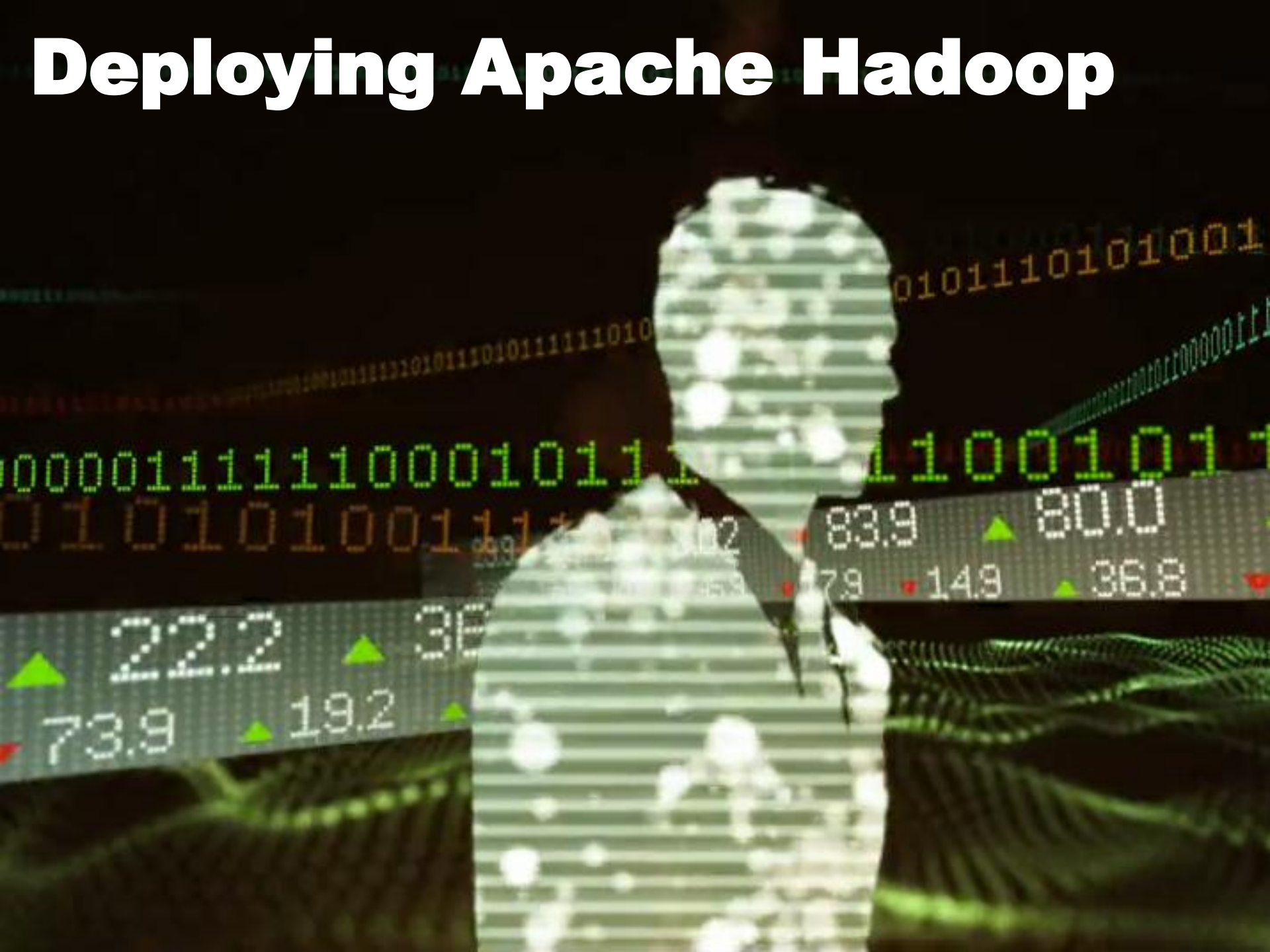
Map Reduce

- Splits a task across processors “near” the data & assembles results
- Self-Healing, High Bandwidth Clustered Storage

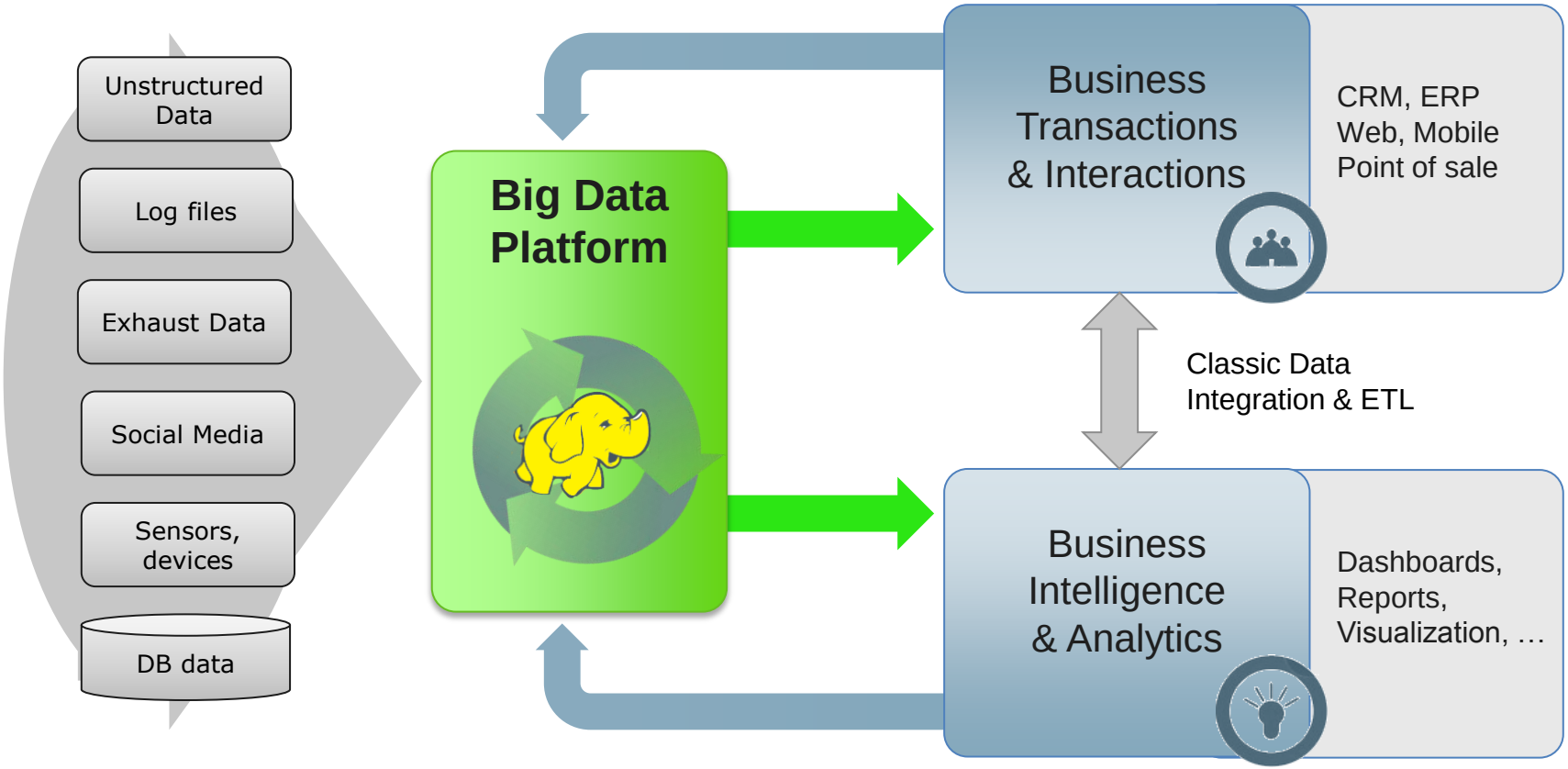
Key Characteristics

- **Scalable**
  - Efficiently store and process petabytes of data
  - Linear scale driven by additional processing and storage
- **Reliable**
  - Redundant storage
  - Failover across nodes and racks
- **Flexible**
  - Store all types of data in any format
  - Apply schema on analysis and sharing of the data
- **Economical**
  - Use commodity hardware
  - Open source software guards against vendor lock-in

Deploying Apache Hadoop



Hadoop in Enterprise Big Data Flows

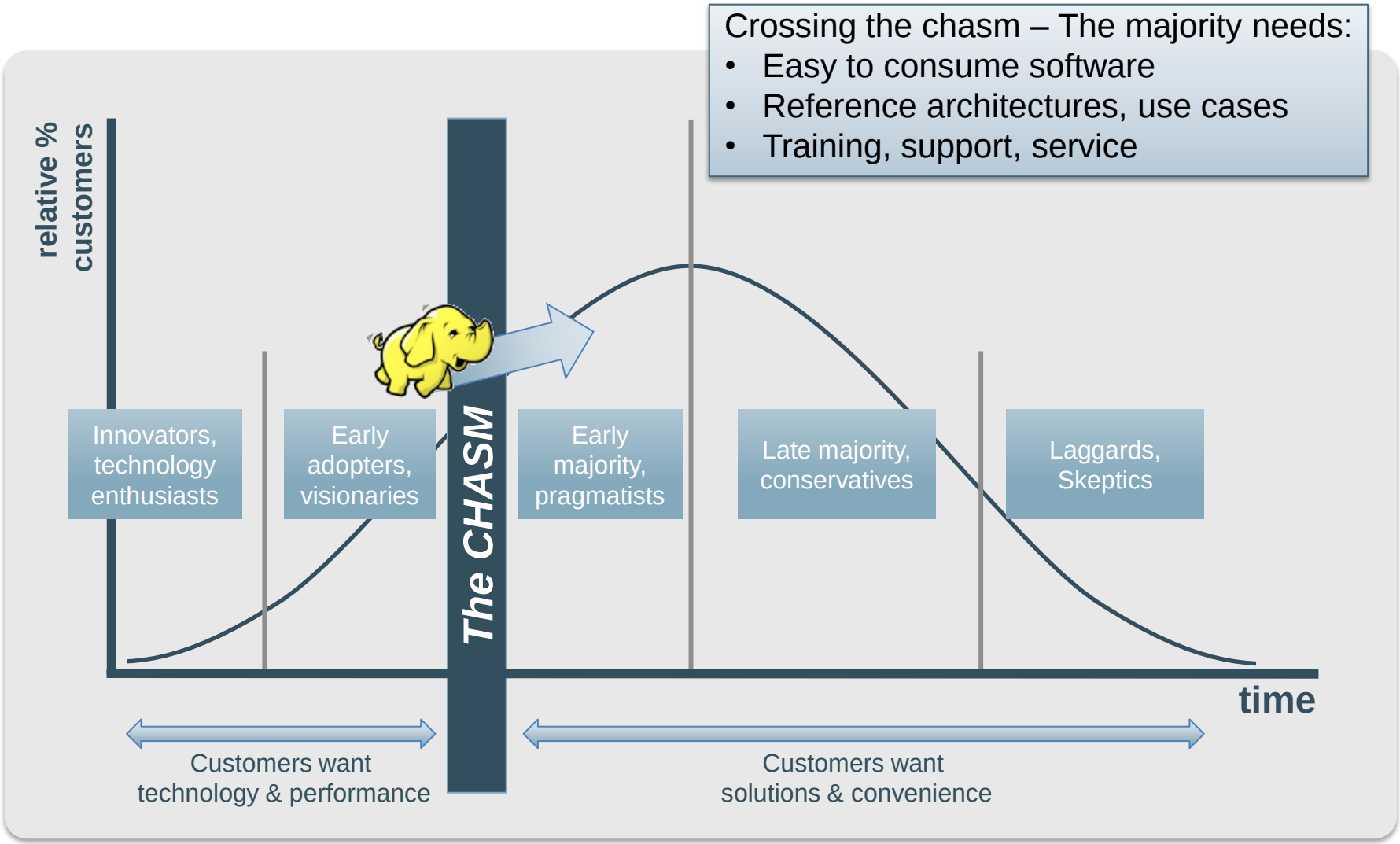


1 Capture Big Data
Collect data from all sources structured &unstructured

2 Process
Transform, refine, aggregate, analyze, report

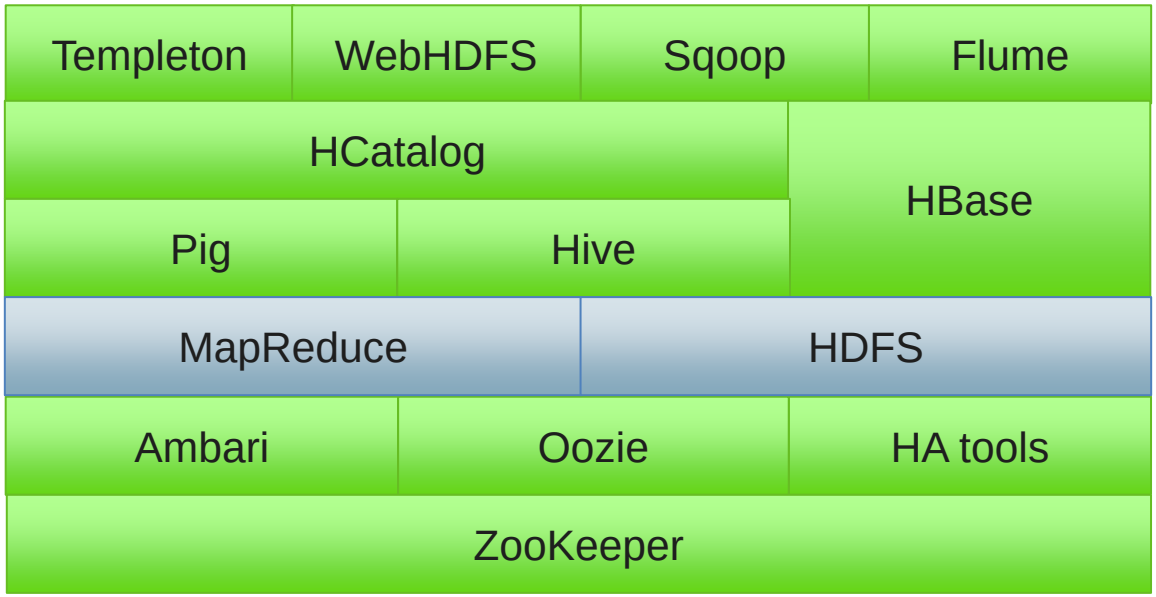
3 Exchange Results
Interoperate and share data with applications/analytics

Hadoop: Poised for Rapid Growth



Source: Geoffrey Moore - Crossing the Chasm

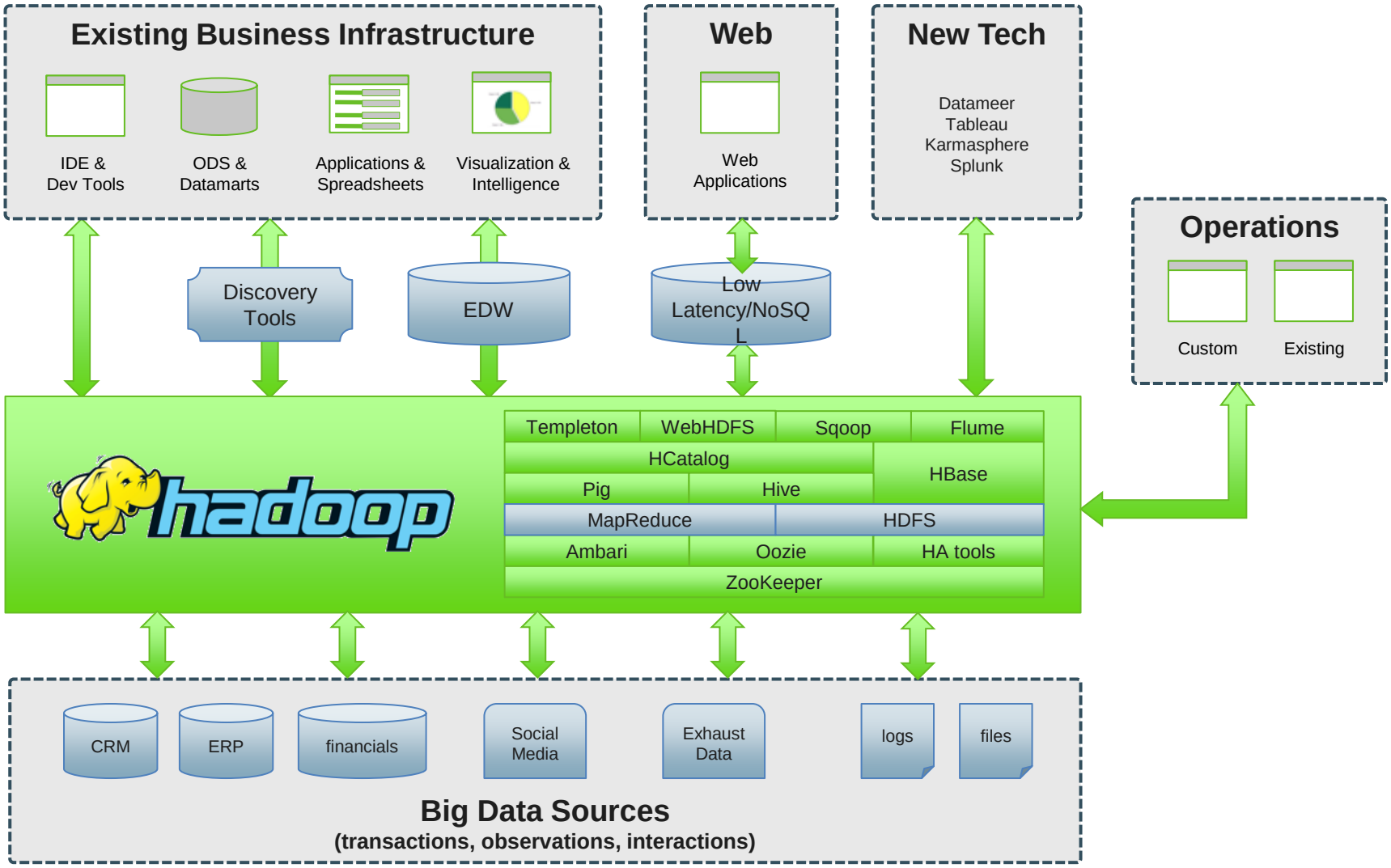
What is a Hadoop “Distribution”



A collection of Apache Hadoop and related technologies to make it easy to install and operate a Hadoop Big Data Platform

- Solves the tricky problem of choosing the right versions of many components that all have different release cycles
- Tested and pre-packaged to ease installation and usage

Hadoop in Enterprise Data Architectures



Hortonworks and Hadoop in the Enterprise

- ***"We believe that by the end of 2015, more than half the world's data will be processed by Apache Hadoop."***

- Hortonworks mission:
  - Provide a complete and 100% open source Apache Hadoop Distribution
  - Invest in Apache Hadoop to make it ***"The enterprise big data platform"***

Distribution

- **Hortonworks Data Platform (HDP)**
- 100% open source

- From the team that leads the Apache release of Hadoop
- Tested, reliable, easy
- Certified w our partners

Ecosystem

- Training and certification programs available

- Web content
- Public classes
- Onsite training
- Direct & via partners

Support

- HDP support available for all phases of project lifecycle
- Hadoop training and certification by the Hadoop experts(web, public, private)
- Professional services team

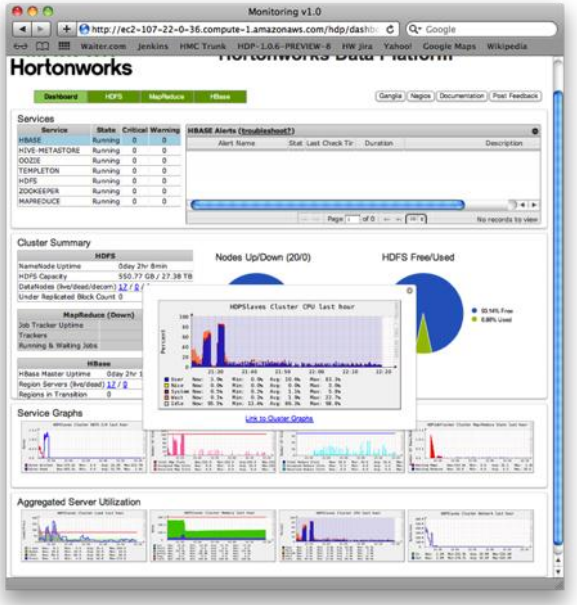
Hortonworks

The Road Ahead

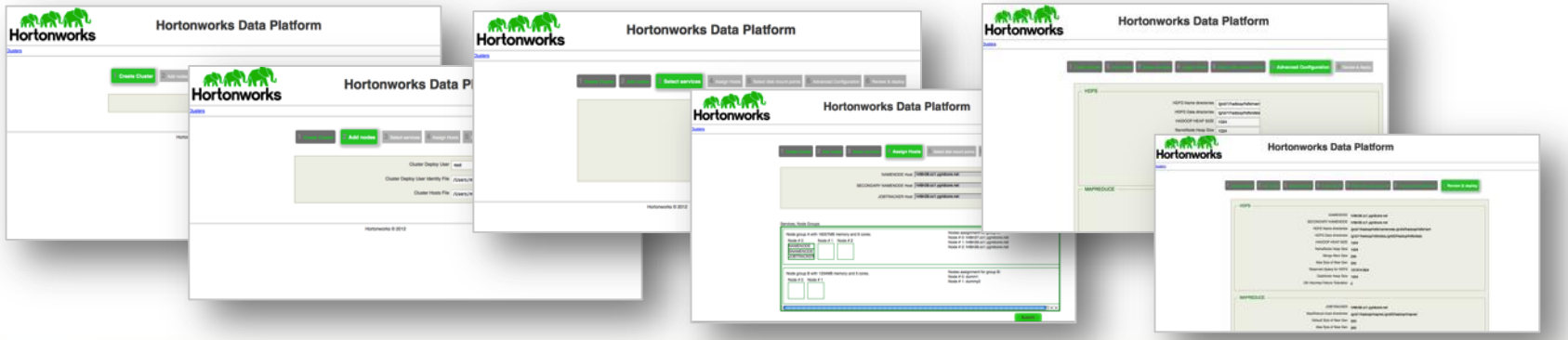
Coming improvements
to Hadoop

Ambari: Management & Monitoring Services

- **Powerful monitoring and alerting dashboards**
  - View topology, health & utilization of cluster
  - Detailed view of cluster operations, server & storage utilization, job status, and performance levels
  - Get alerts to critical events



- **Simple installation & provisioning**
  - Easy configuration process
  - One-click deployment for clusters of all sizes
  - Analyzes/recommends optimal services configuration
  - Automatically configures mount points in the cluster

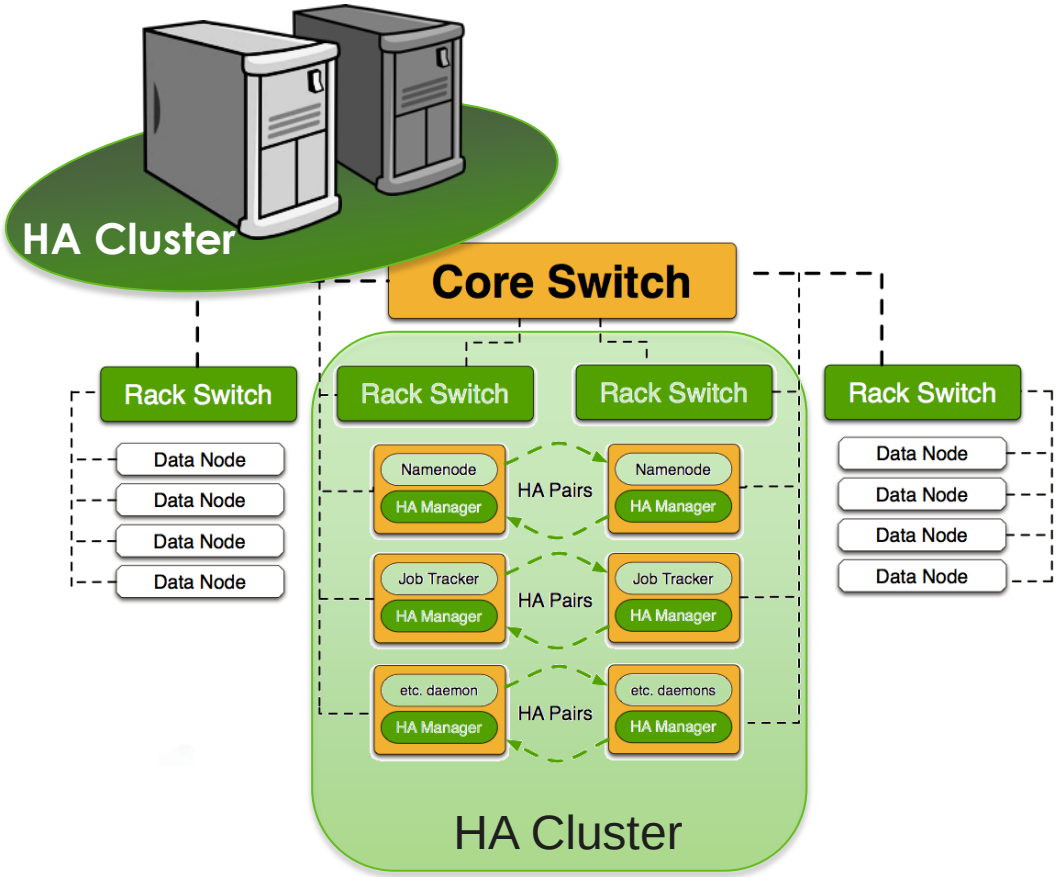


Hive + HCatalog

- **ODBC / JDBC support**
  - Support for BI tools such as Tableau & Excel
- **HCatalog**
  - Allow the use of Hive tables with all Hadoop tools (Pig, MapReduce…)
  - Allow use of Hadoop data with MPP DBs (Aster SQLH)
  - Providing a new higher level of abstraction for Hadoop data
- **Performance**
  - New column oriented file formats
  - Lots of inner loop performance improvements
  - Improved query planning
  - Yarn integration

Full Stack High Availability

Proven HA solutions with proven Hadoop 1.0 & 2.0



Failover and restart for

- NameNode
- JobTracker
- Other services to come…

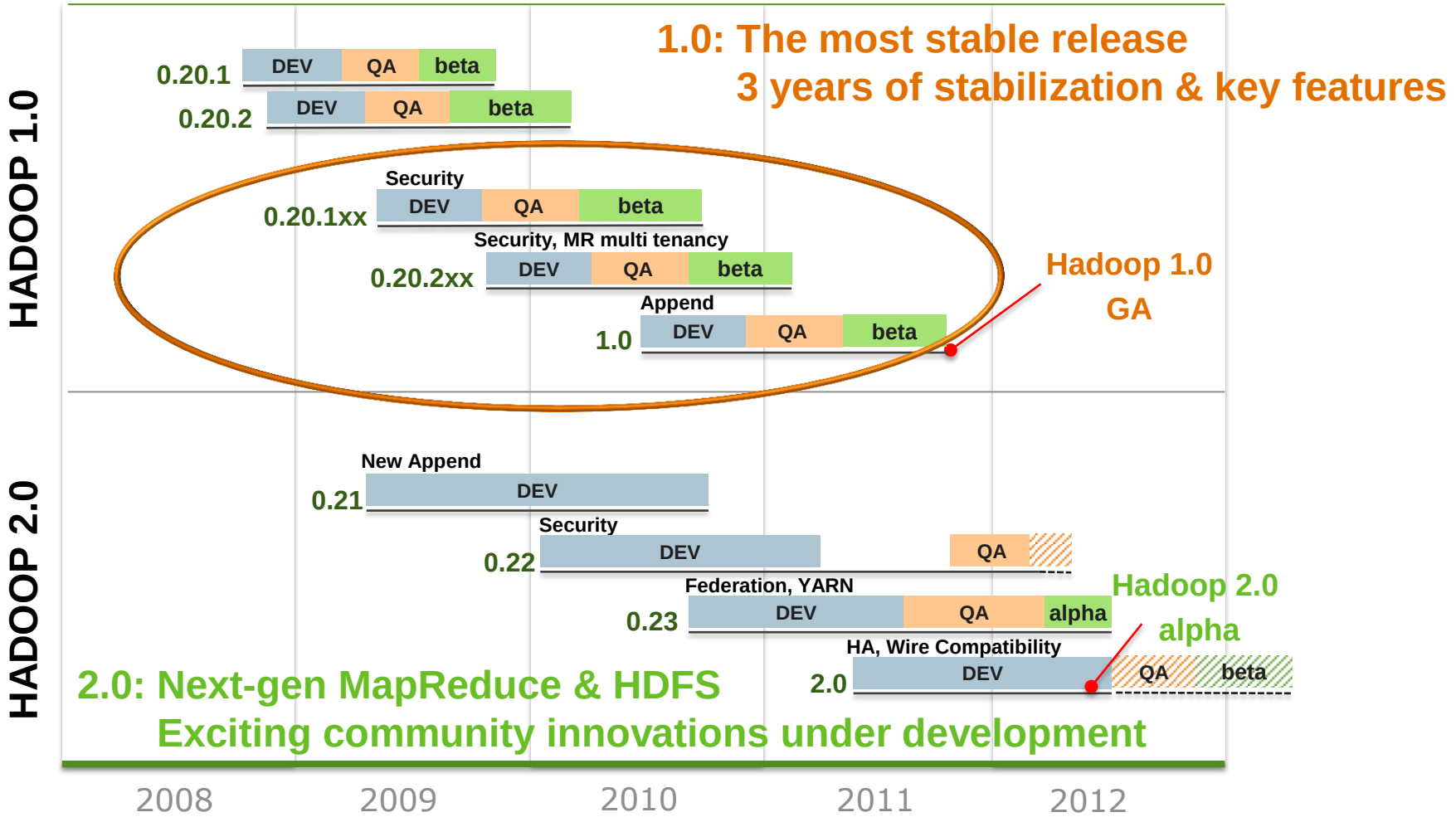
Open API allows use of Proven HA from multiple vendors

Minimized changes to clients and configuration

Auto-detects failures:

- Services, OS & Hardware

Timeline: Apache Hadoop 1.0 & 2.0



Hadoop 2.0 Innovations

- **Focus on Scale and Community Innovation**
  - Designed to support 10,000+ computer clusters
  - Extensible design to encourage innovation

- **YARN: Next Generation Execution**
  - Improves MapReduce performance
  - Supports **new frameworks** beyond MapReduce!
    - Low latency, Streaming, Services
    - Do more with a single Hadoop cluster

- **HDFS 2.0**
  - Federation: Isolation & extensibility via multiple NameNodes
  - Checkpoint support
  - Performance improvement

Thank You!

Questions & Answers

Follow: @hortonworks