Hortonworks

Big Business Value from Big Data and Hadoop

Topics

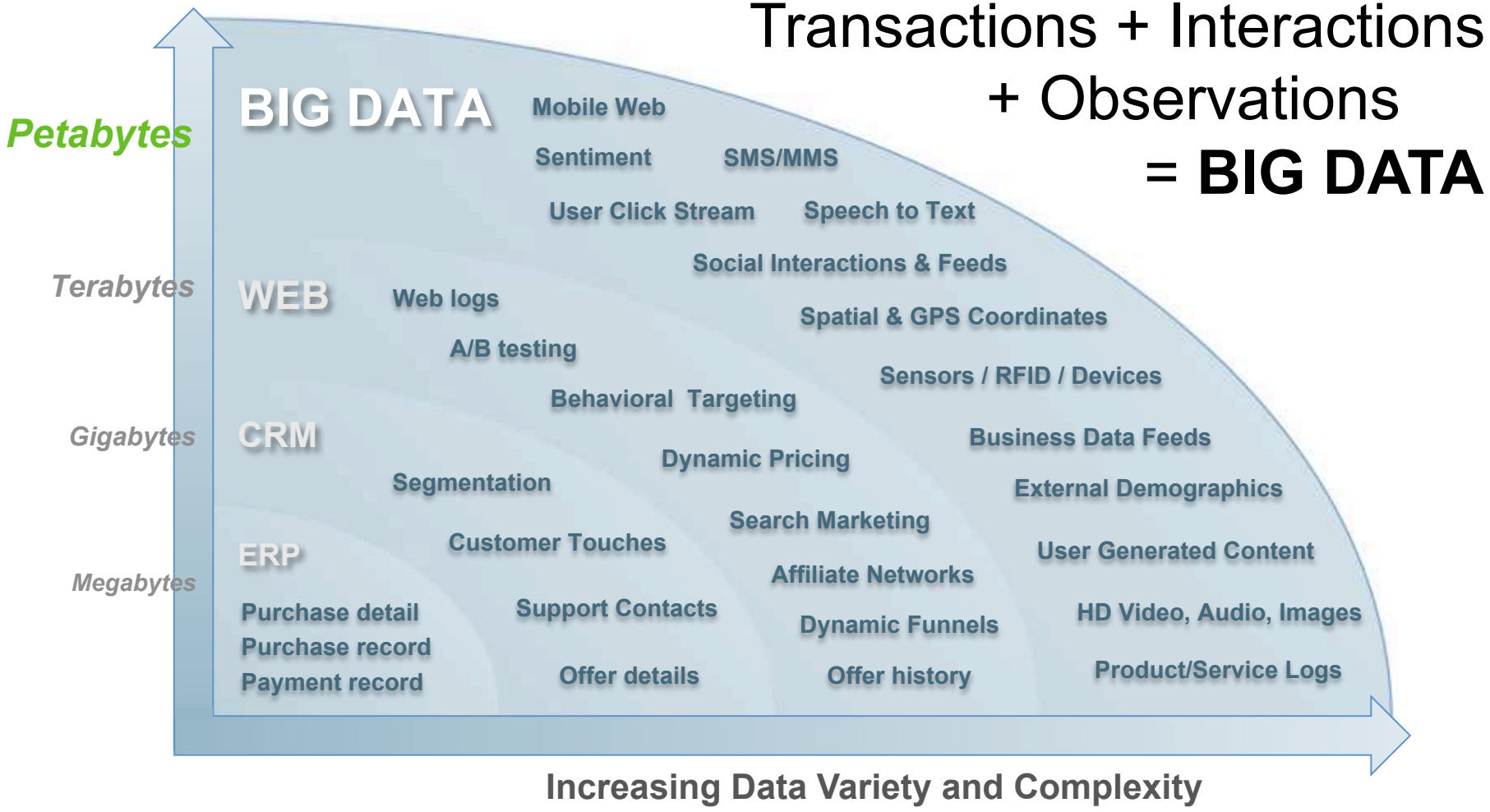
- **The Big Data Explosion: Hype or Reality**
- **Introduction to Apache Hadoop**
- **The Business Case for Big Data**
- **Hortonworks Overview & Product Demo**

Hortonworks

Big Data: Hype or Reality?

What is Big Data?

Big Data: Changing The Game for Organizations

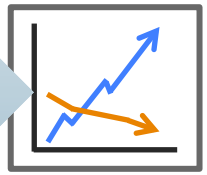


Next Generation Data Platform Drivers

Organizations will need to become more data driven to compete

Business Drivers

- Enable new business models & drive faster growth (20%+)
- Find insights for competitive advantage & optimal returns



Technical Drivers

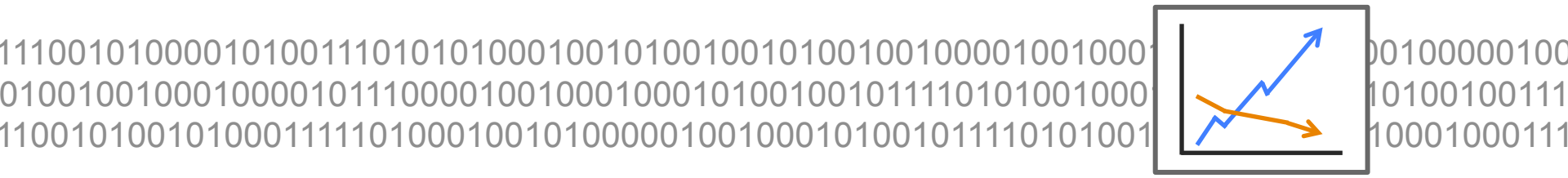
- Data continues to grow exponentially
- Data is increasingly everywhere and in many formats
- Legacy solutions unfit for new requirements growth

Financial Drivers

- Cost of data systems, as % of IT spend, continues to grow
- Cost advantages of commodity hardware & open source

What is a Data Driven Business?

- **DEFINITION**
  Better use of available data in the decision making process
- **RULE**
  Key metrics derived from data should be tied to goals
- **PROVEN RESULTS**
  Firms that adopt Data-Driven Decision Making have output and productivity that is *5-6% higher* than what would be expected given their investments and usage of information technology*



* "Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?" *Brynjolfsson, Hitt and Kim (April 22, 2011)*

Path to Becoming Big Data Driven

4

Key Considerations for a Data Driven Business

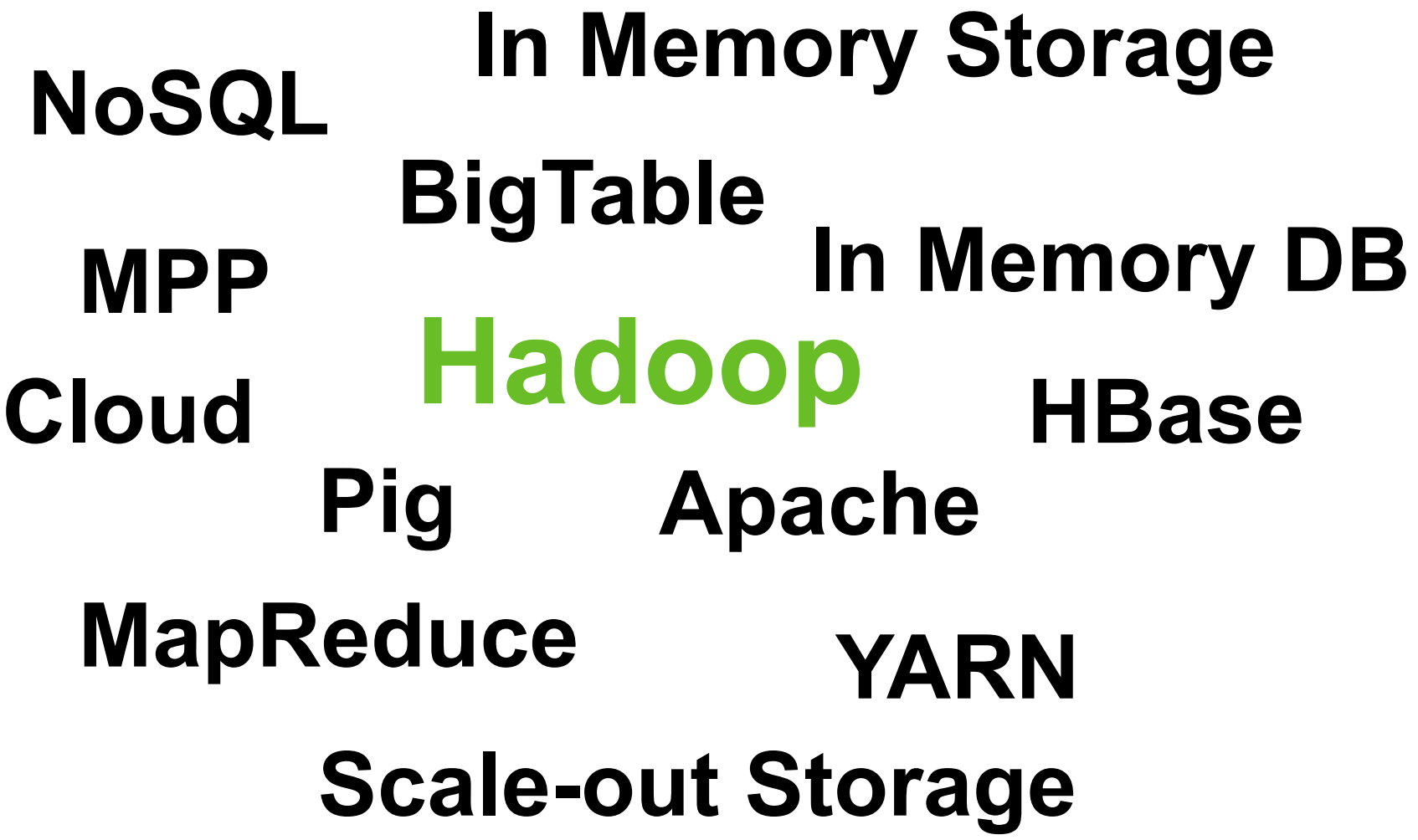
1. Large web properties were born this way, you may have to adapt a strategy
2. Start with a project tied to a key objective or KPI – Don't OVER engineer
3. Make sure your Big Data strategy "fits" your organization and grow it over time
4. Don't do big data just to do big data – you can get lost in all that data

"Simply put, because of big data, managers can measure, and hence know, radically more about their businesses, and directly translate that knowledge into improved decision making & performance."

- Erik Brynjolfsson and Andrew McAfee

Harvard Business Review

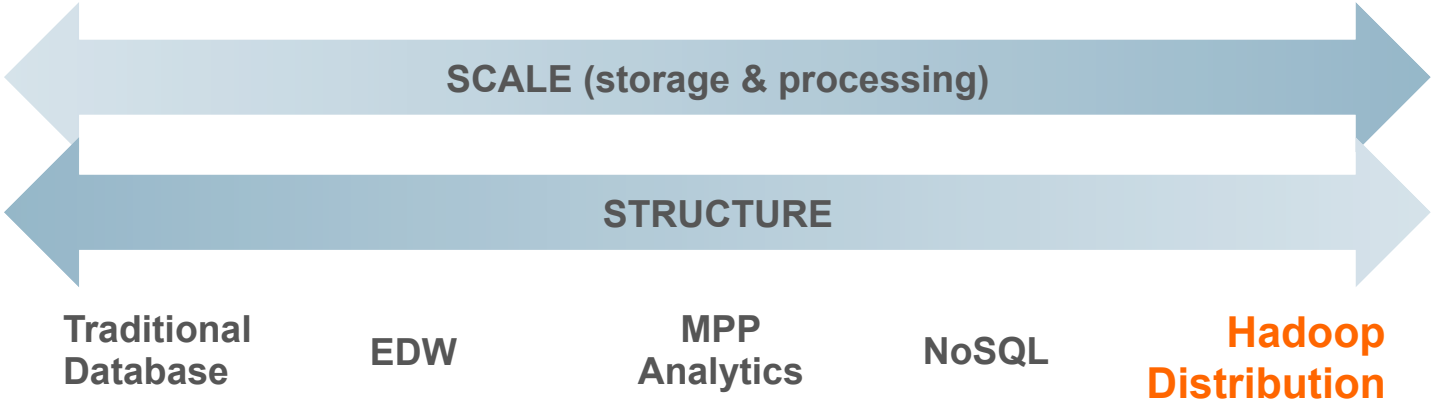
Wide Range of New Technologies



A Few Definitions of New Technologies

- **NoSQL**
  - A broad range of new database management technologies mostly focused on low latency access to large amounts of data.
- **MPP or "in database analytics"**
  - Widely used approach in data warehouses that incorporates analytic algorithms with the data and then distribute processing
- **Cloud**
  - The use of computing resources that are delivered as a service over a network
- **Hadoop**
  - Open source data management software that combines massively parallel computing and highly-scalable distributed storage

Big Data: It's About Scale & Structure

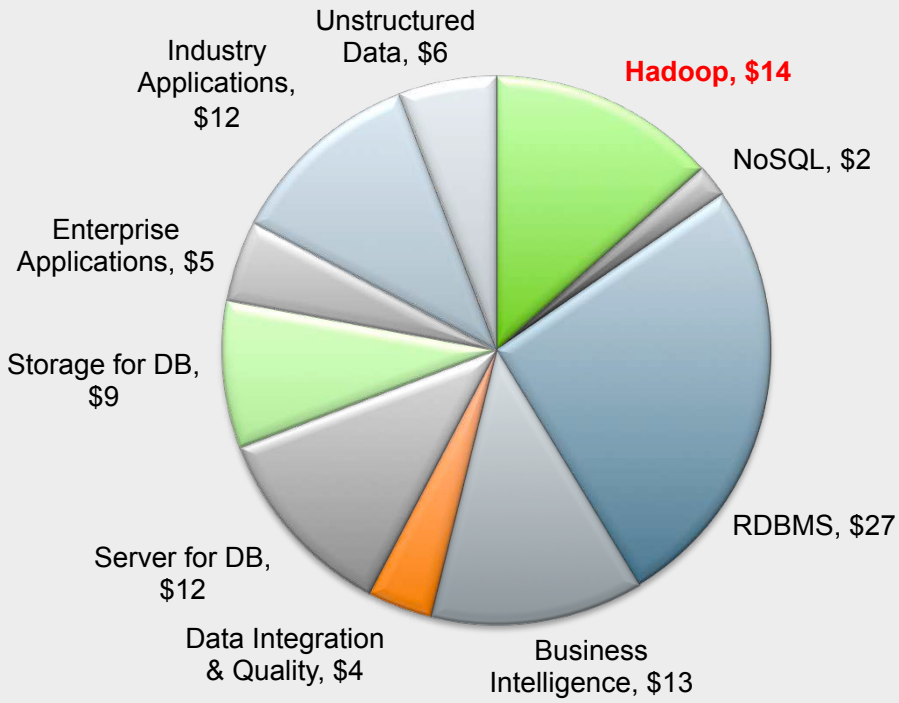


Standards and structured	governance	Loosely structured
Limited, no data processing	processing	Processing coupled w/ data
Structured	data types	Multi and unstructured

A JDBC connector to Hive does not make you "big data"

Hadoop Market: Growing & Evolving

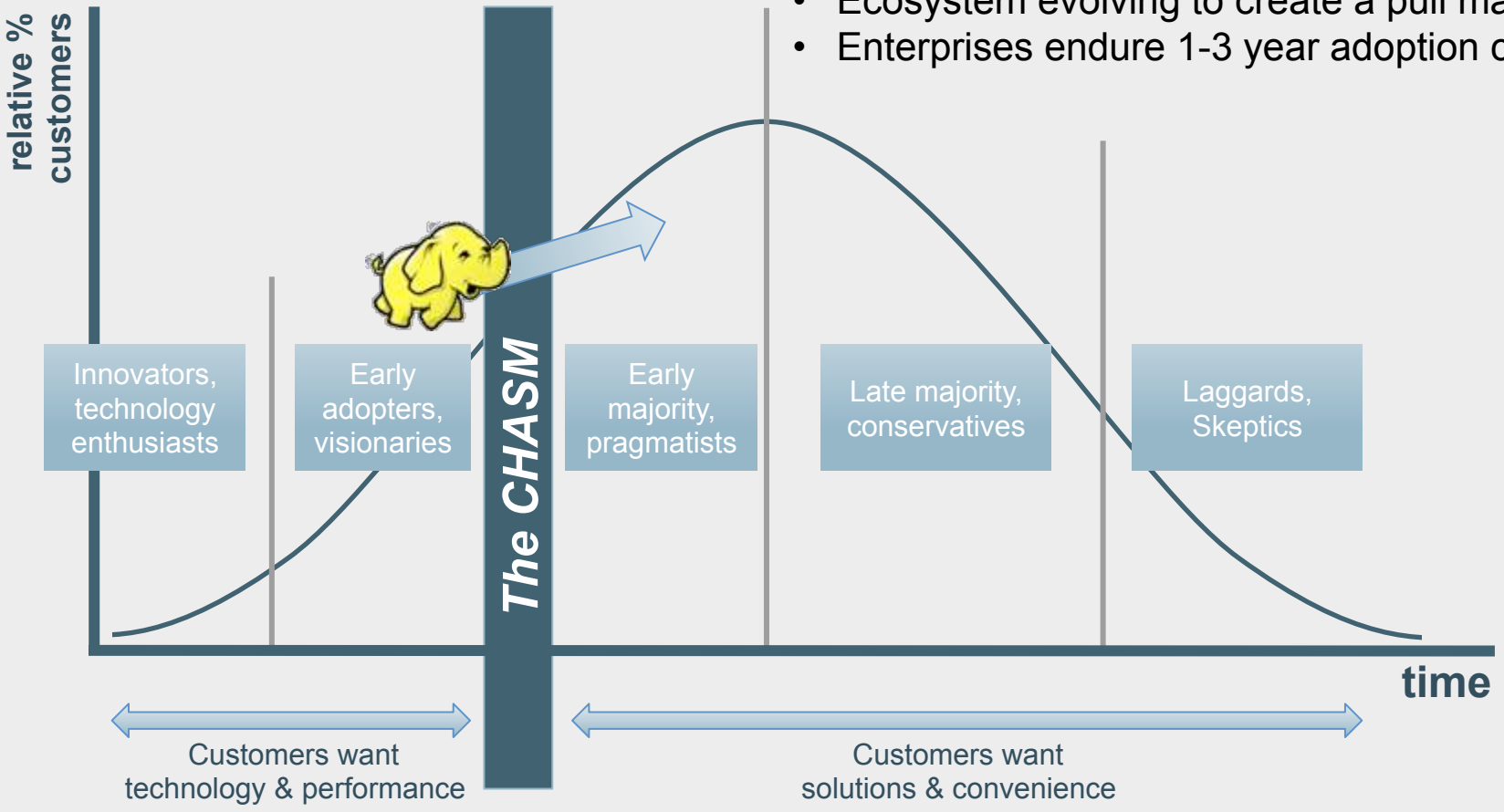
- **Big data outranks virtualization as #1 trend driving spending initiatives**
  - Barclays CIO survey April 2012
- **Overall market at $100B**
  - Hadoop 2nd only to RDBMS in potential
- **Estimates put market growth > 40% CAGR**
  - IDC expects Big Data tech and services market to grow to $16.9B in 2015
  - According to JPMC 50% of Big Data market will be influenced by Hadoop



Source: Big Data II – New Themes, New Opportunities, BofA Merrill Lynch, March 2012

Hadoop: Poised for Rapid Growth

- Orgs looking for use cases & ref arch
- Ecosystem evolving to create a pull market
- Enterprises endure 1-3 year adoption cycle



Source: Geoffrey Moore - Crossing the Chasm

What's Needed to Drive Success?

- **Enterprise tooling to become a complete data platform**
  - Open deployment & provisioning
  - Higher quality data loading
  - Monitoring and management
  - APIs for easy integration

www.hortonworks.com/moore

- **Ecosystem needs support & development**
  - Existing infrastructure vendors need to continue to integrate
  - Apps need to continue to be developed on this infrastructure
  - Well defined use cases and solution architectures need to be promoted

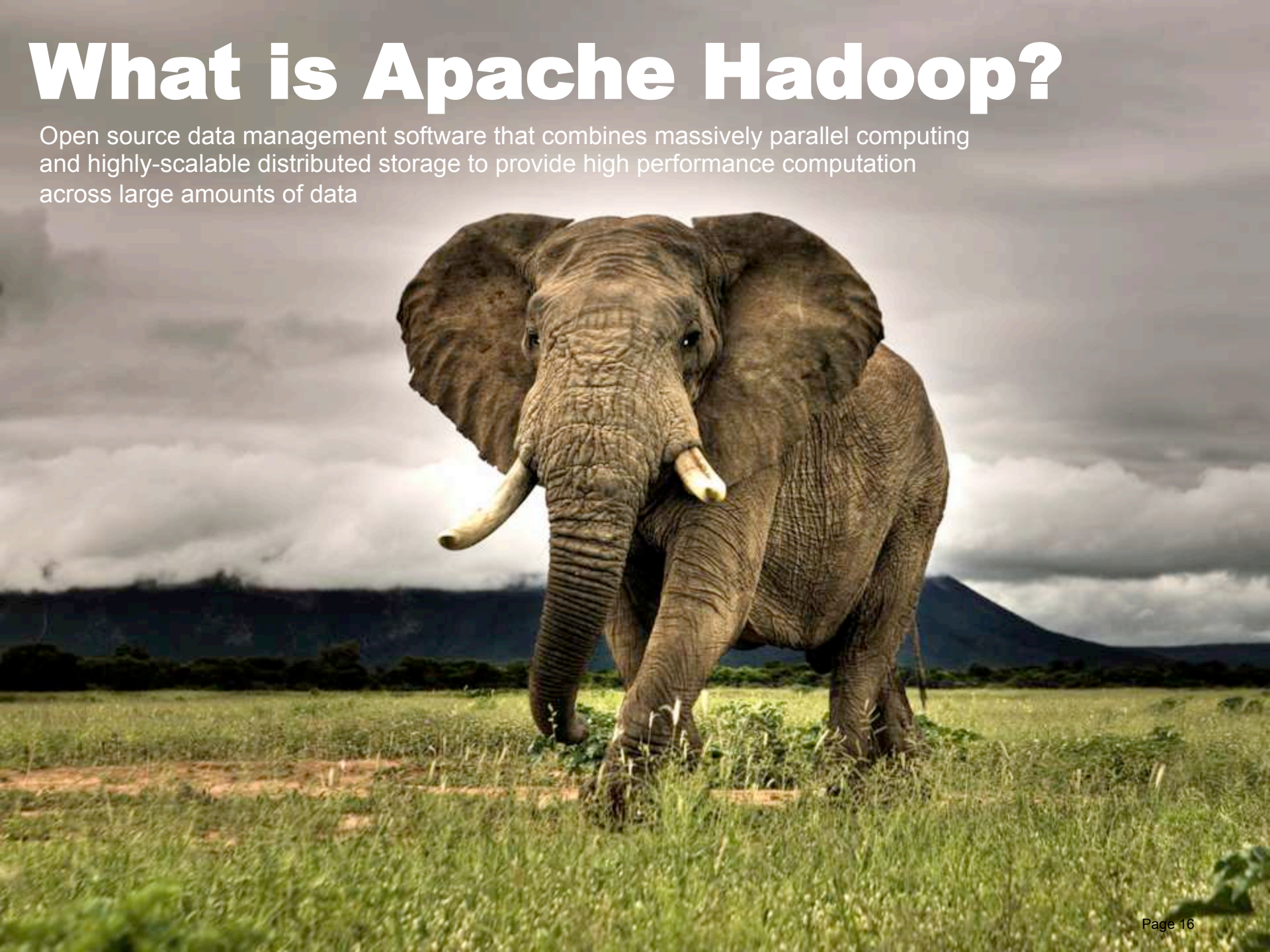
- **Market needs to rally around core Apache Hadoop**
  - To avoid splintering/market distraction
  - To accelerate adoption

Hortonworks

What is Hadoop?

What is Apache Hadoop?

Open source data management software that combines massively parallel computing and highly-scalable distributed storage to provide high performance computation across large amounts of data

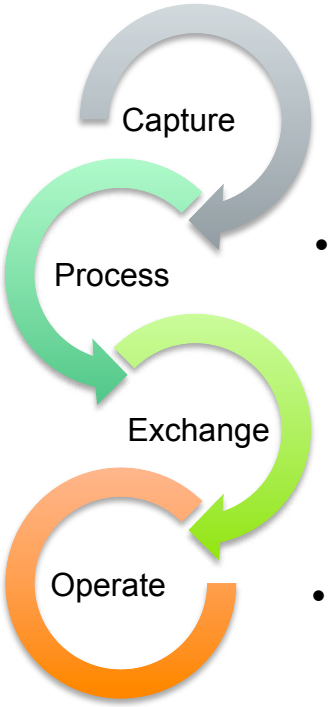


Big Data Driven Means a New Data Platform

- **Facilitate data capture**
  - Collect data from all sources - structured and unstructured data
  - At all speeds batch, asynchronous, streaming, real-time
- **Comprehensive data processing**
  - Transform, refine, aggregate, analyze, report
- **Simplified data exchange**
  - Interoperate with enterprise data systems for query and processing
  - Share data with analytic applications
  - Data between analyst
- **Easy to operate platform**
  - Provision, monitor, diagnose, manage at scale
  - Reliability, availability, affordability, scalability, interoperability

Enterprise Data Platform Requirements

A data platform is an integrated set of components that allows you to capture, process and share data in any format, at scale



- Store and integrate data sources in real time and batch
- Process and manage data of any size and includes tools to sort, filter, summarize and apply basic functions to the data
- Open and promote the exchange of data with new & existing external applications
- Includes tools to manage and operate and gain insight into performance and assure availability

Apache Hadoop

Open Source data management with scale-out storage & processing

Processing

MapReduce

- Splits a task across processors "near" the data & assembles results
- Distributed across "nodes"

Storage

HDFS

- Self-healing, high bandwidth clustered storage
- Natively redundant
- NameNode tracks locations

Apache Hadoop Characteristics

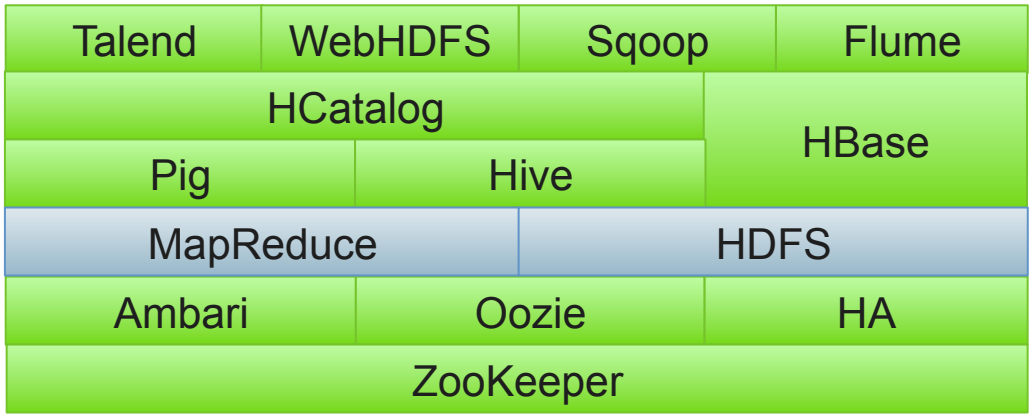
- **Scalable**
  - Efficiently store and process petabytes of data
  - Linear scale driven by additional processing and storage
- **Reliable**
  - Redundant storage
  - Failover across nodes and racks
- **Flexible**
  - Store all types of data in any format
  - Apply schema on analysis and sharing of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

Tale of the Tape

Relational	VS.	Hadoop
Required on write	schema	Required on read
Reads are fast	speed	Writes are fast
Standards and structured	governance	Loosely structured
Limited, no data processing	processing	Processing coupled with data
Structured	data types	Multi and unstructured
Interactive OLAP Analytics Complex ACID Transactions Operational Data Store	best fit use	Data Discovery Processing unstructured data Massive Storage/Processing

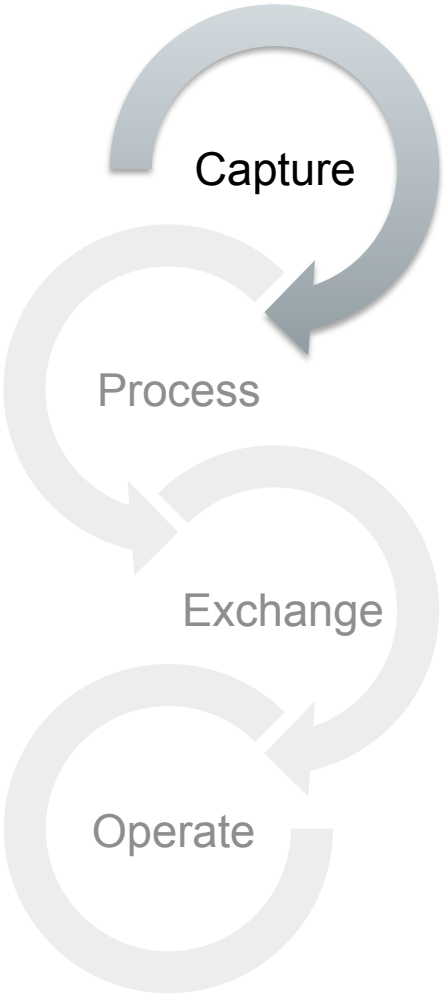
What is a Hadoop “Distribution”

A complimentary set of open source technologies that make up a complete data platform



- Tested and pre-packaged to ease installation and usage
- Collects the right versions of the components that all have different release cycles and ensures they work together

Apache Hadoop Related Projects



Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

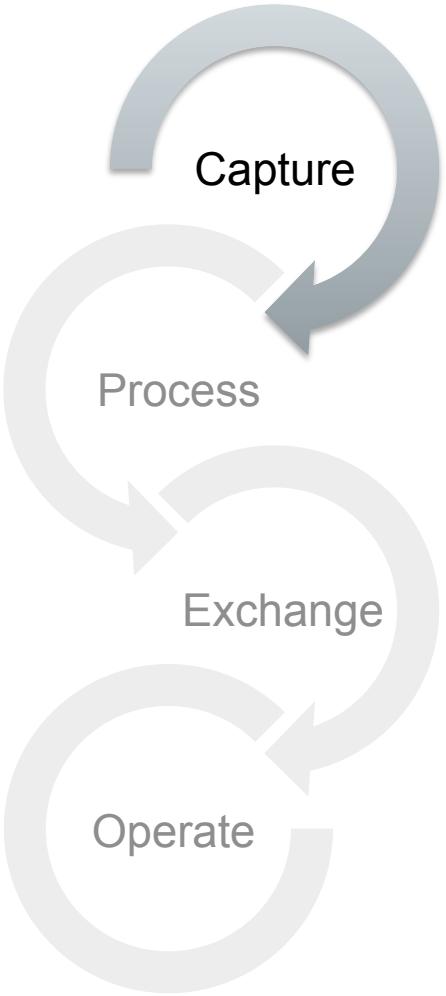
WebHDFS

- REST API that supports the complete FileSystem interface for HDFS.
- Move data in and out and delete from HDFS
- Perform file and directory functions

- webhdfs://<HOST>:<HTTP PORT>/PATH

- Standard and included in version 1.0 of Hadoop

Apache Hadoop Related Projects

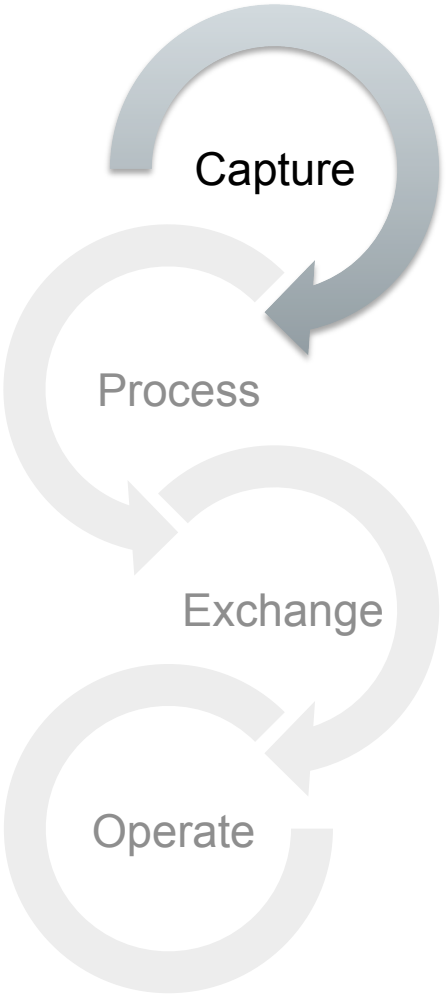


Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Sqoop

- Sqoop is a set of tools that allow non-Hadoop data stores to interact with traditional relational databases and data warehouses.
- A series of connectors have been created to support explicit sources such as Oracle & Teradata
- It moves data in and out of Hadoop
- SQ-OOP: SQL to Hadoop

Apache Hadoop Related Projects



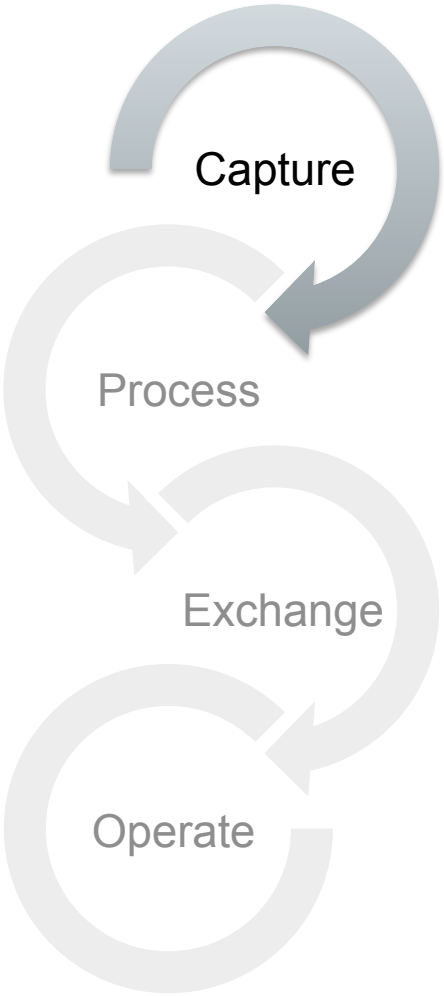
Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Flume

- Distributed service for efficiently collecting, aggregating, and moving streams of log data into HDFS
- Streaming capability with many failover and recovery mechanisms

- Often used to move web log files directly into Hadoop

Apache Hadoop Related Projects

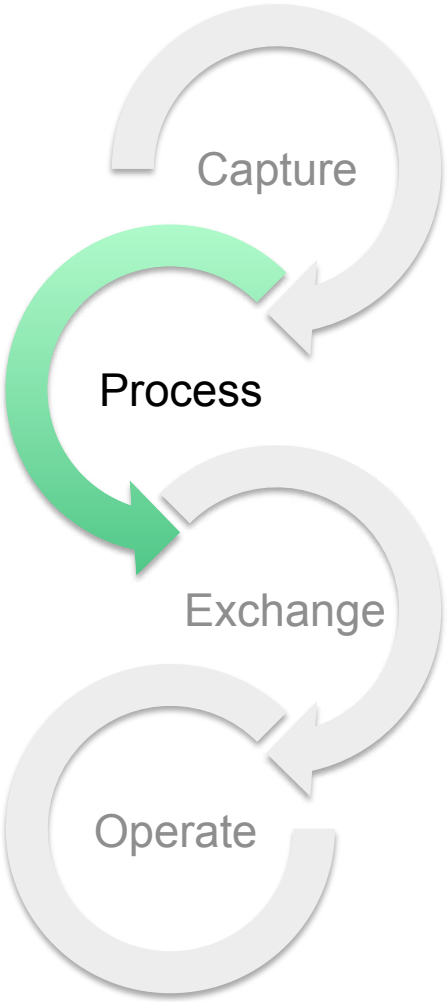


Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache HBase

- HBase is a non-relational database. It is columnar and provides fault-tolerant storage and quick access to large quantities of sparse data. It also adds transactional capabilities to Hadoop, allowing users to conduct updates, inserts and deletes.

Apache Hadoop Related Projects

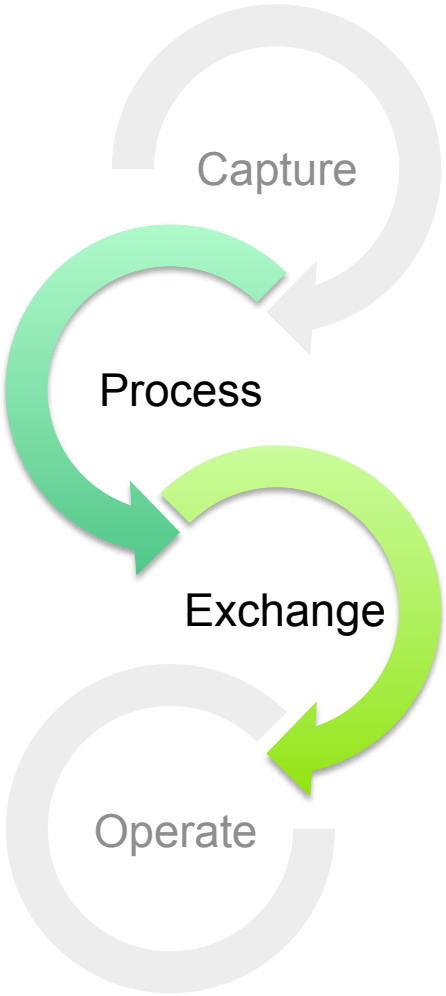


Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Pig

- Apache Pig allows you to write complex map reduce transformations using a simple scripting language. Pig latin (the language) defines a set of transformations on a data set such as aggregate, join and sort among others. Pig Latin is sometimes extended using UDF (User Defined Functions), which the user can write in Java and then call directly from the language.

Apache Hadoop Related Projects

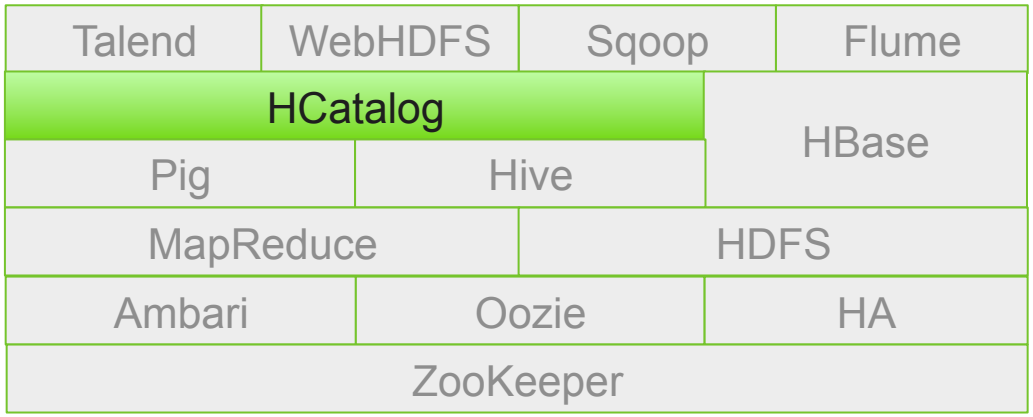


Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Hive

- Apache Hive is a data warehouse infrastructure built on top of Hadoop (originally by Facebook) for providing data summarization, ad-hoc query, and analysis of large datasets. It provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL (HQL).

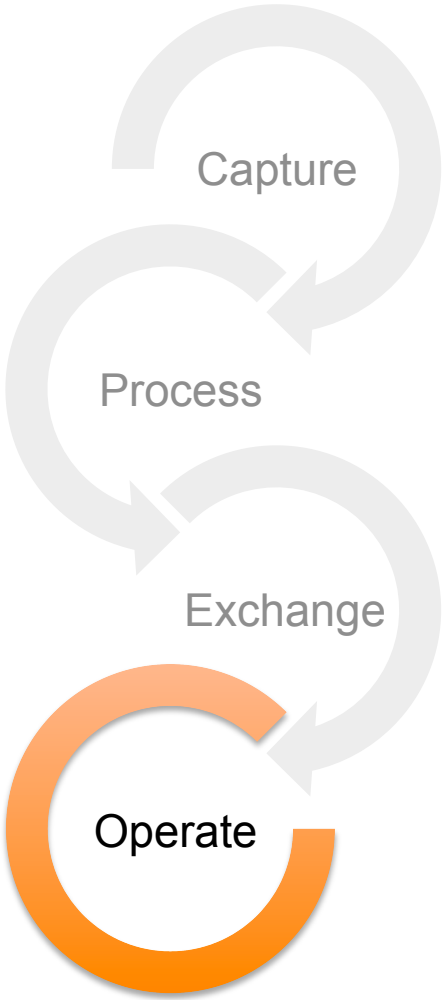
Apache Hadoop Related Projects



Apache HCatalog

- HCatalog is a metadata management service for Apache Hadoop. It opens up the platform and allows interoperability across data processing tools such as Pig, Map Reduce and Hive. It also provides a table abstraction so that users need not be concerned with where or how their data is stored.

Apache Hadoop Related Projects



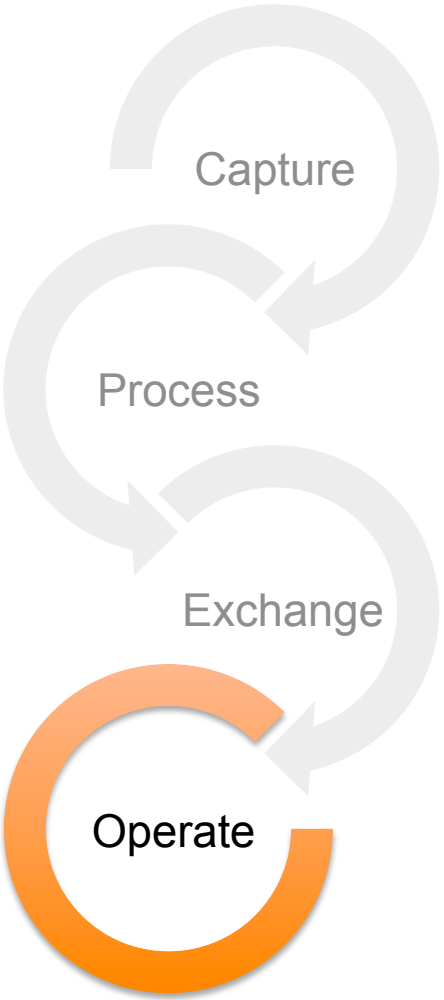
Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Ambari

- Ambari is a monitoring, administration and lifecycle management project for Apache Hadoop clusters
- It provides a mechanism to provisions nodes

- Operationalizes Hadoop.

Apache Hadoop Related Projects

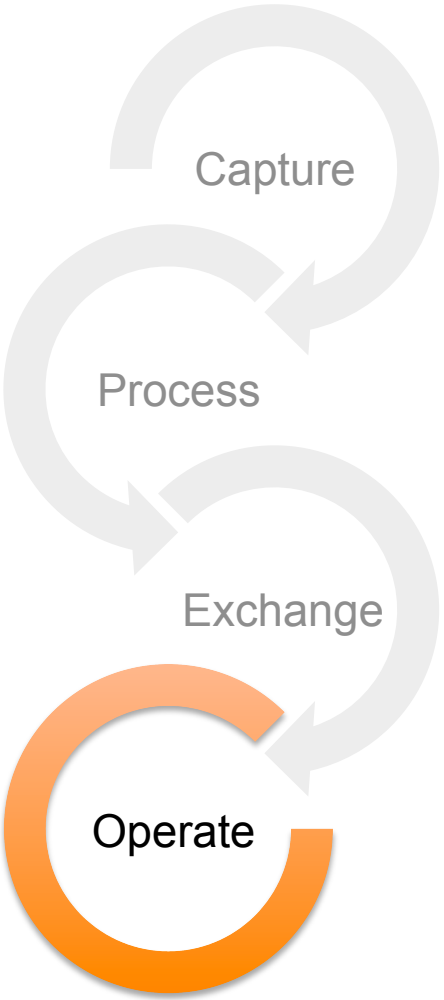


Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

Apache Oozie

- Oozie coordinates jobs written in multiple languages such as Map Reduce, Pig and Hive. It is a workflow system that links these jobs and allows specification of order and dependencies between them.

Apache Hadoop Related Projects



Talend	WebHDFS	Sqoop	Flume
HCatalog			HBase
Pig	Hive		
MapReduce		HDFS	
Ambari	Oozie	HA	
ZooKeeper			

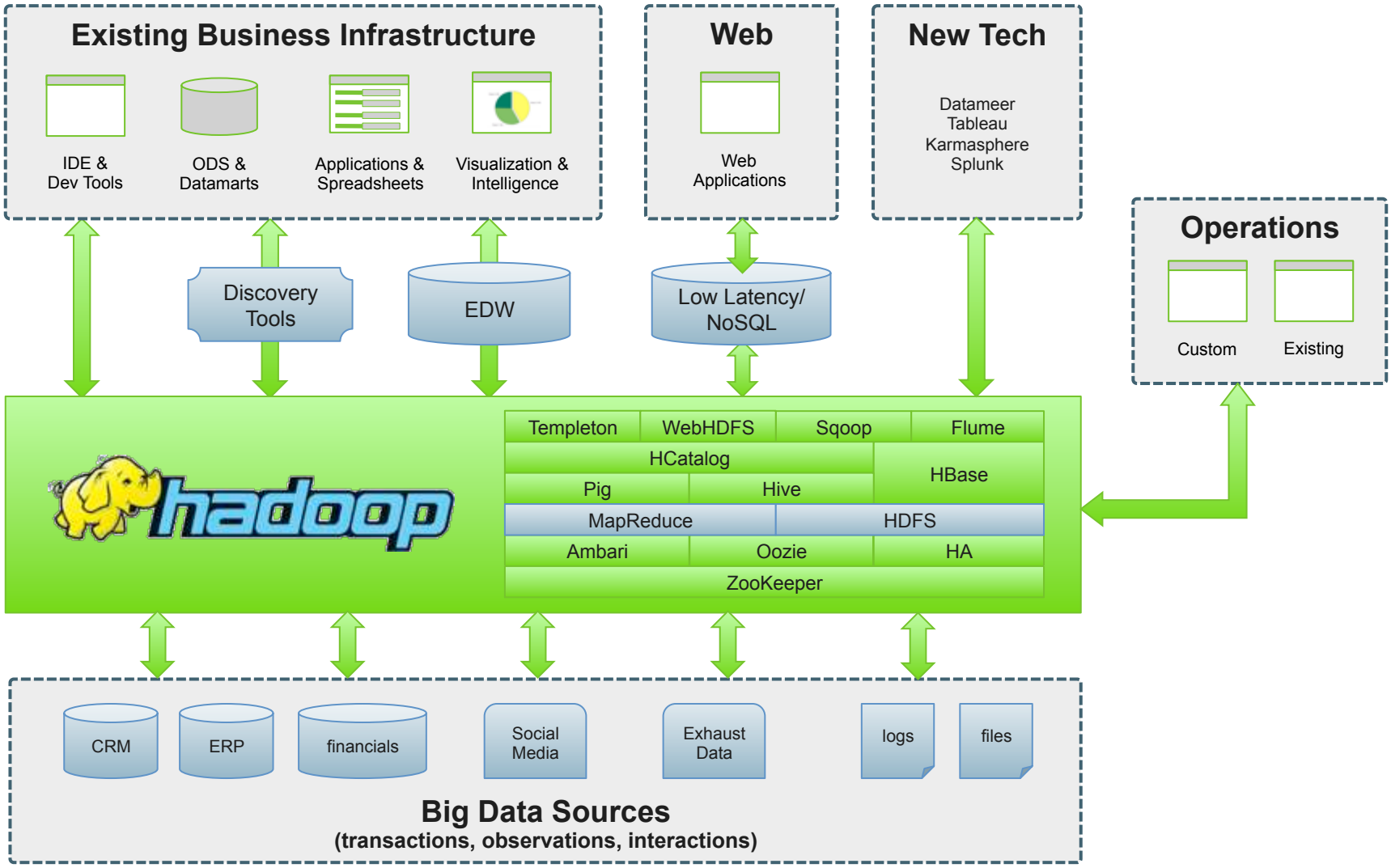
Apache ZooKeeper

- ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

Using Hadoop “out of the box”

1. **Provision your cluster w/ Ambari**
2. **Load your data into HDFS**
   - WebHDFS, distcp, Java
3. **Express queries in Pig and Hive**
   - Shell or scripts
4. **Write some UDFs and MR-jobs by hand**
5. **Graduate to using data scientists**
6. **Scale your cluster and your analytics**
   - Integrate to the rest of your enterprise data silos (TOPIC OF TODAY’S DISCUSSION)

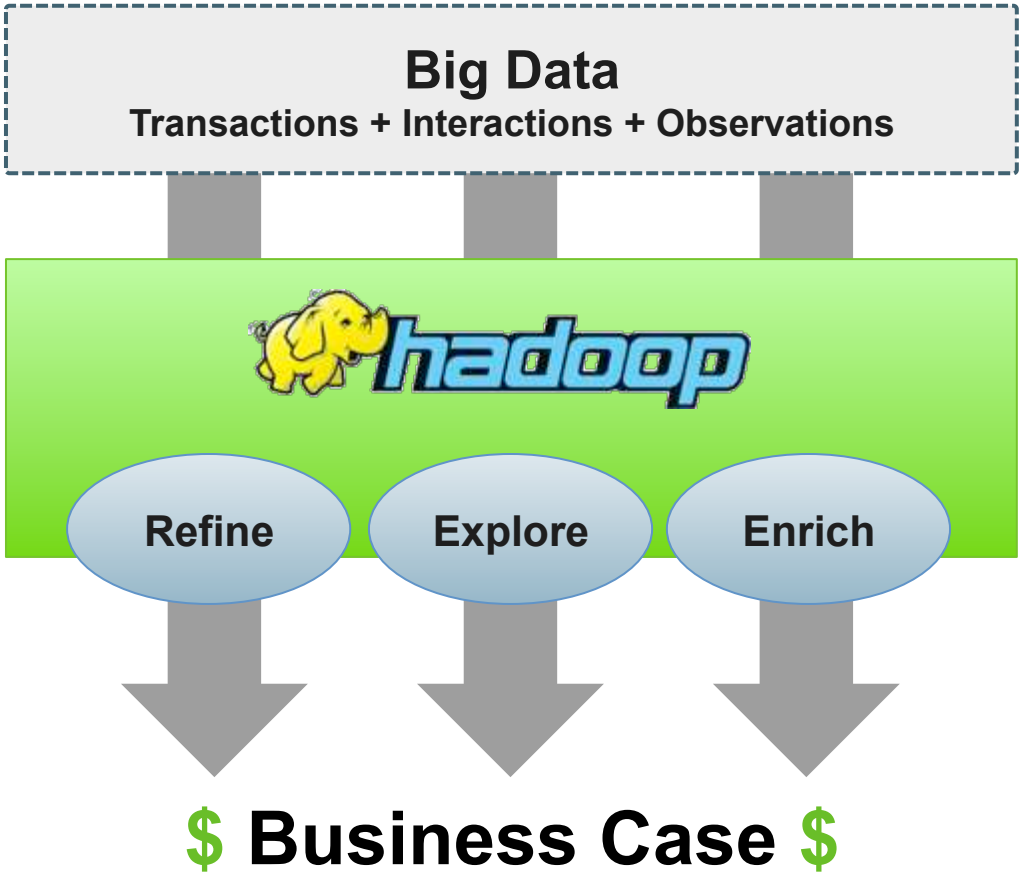
Hadoop in Enterprise Data Architectures



Hortonworks

The Business Case for Hadoop

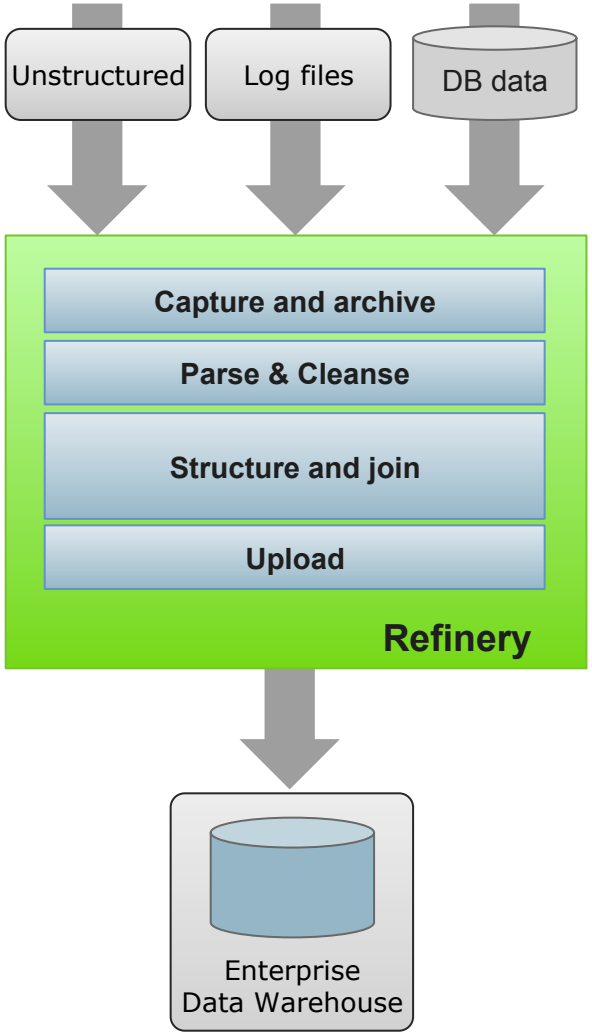
Apache Hadoop: Patterns of Use



Operational Data Refinery

Hadoop as platform for ETL modernization

Refine | Explore | Enrich



Capture

- Capture new unstructured data along with log files all alongside existing sources
- Retain inputs in raw form for audit and continuity purposes

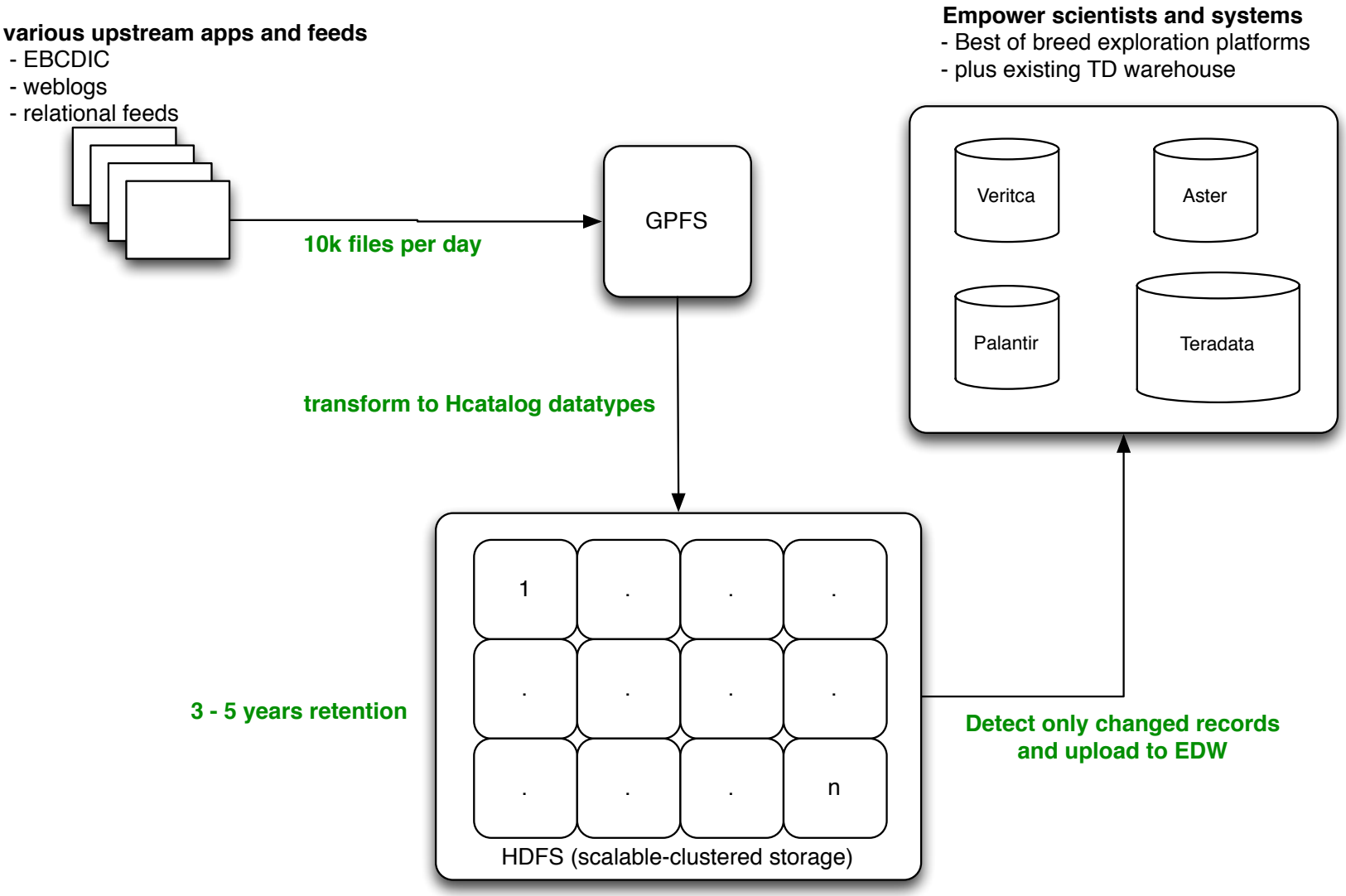
Process

- Parse the data & cleanse
- Apply structure and definition
- Join datasets together across disparate data sources

Exchange

- Push to existing data warehouse for downstream consumption
- Feeds operational reporting and online systems

“Big Bank” Data Refinery Pipeline



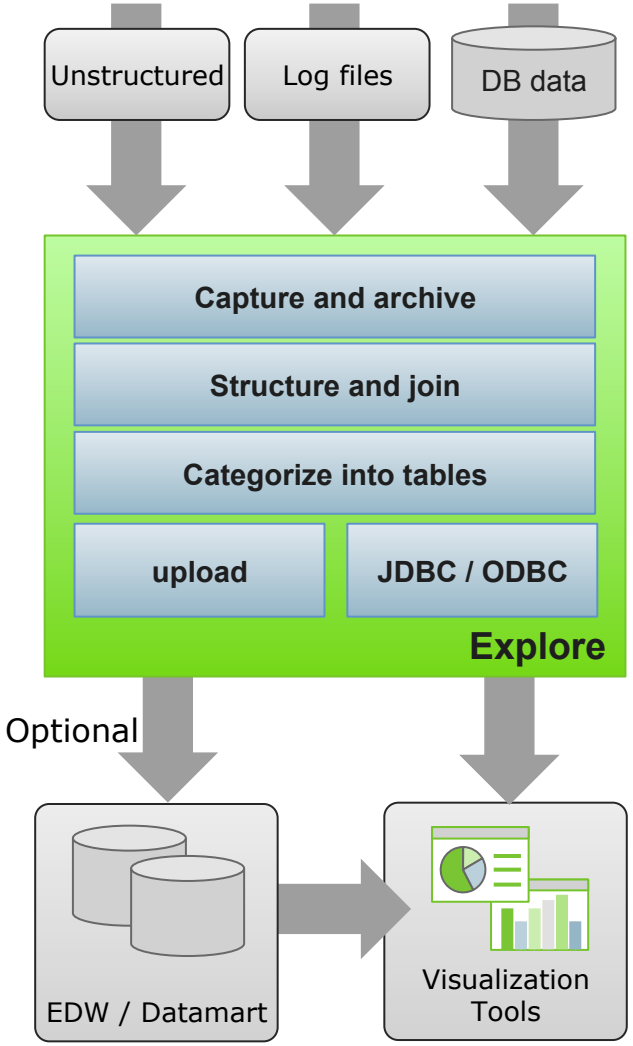
“Big Bank” Data Refinery Key Benefits

- **Capture**
  - Retain 3 – 5 years instead of 2 – 10 days
  - Lower costs
  - Improved compliance
- **Process**
  - Turn upstream raw dumps into small list of “new, update, delete” customer records
  - Convert fixed-width EBCDIC to UTF-8 (Java and DB compatible)
  - Turn raw weblogs into sessions and behaviors
- **Exchange**
  - Insert into Teradata for downstream “as-is” reporting and tools
  - Insert into new exploration platform for scientists to play with

Big Data Exploration & Visualization

Hadoop as agile, ad-hoc data mart

Refine | Explore | Enrich



Capture

- Capture multi-structured data and retain inputs in raw form for iterative analysis

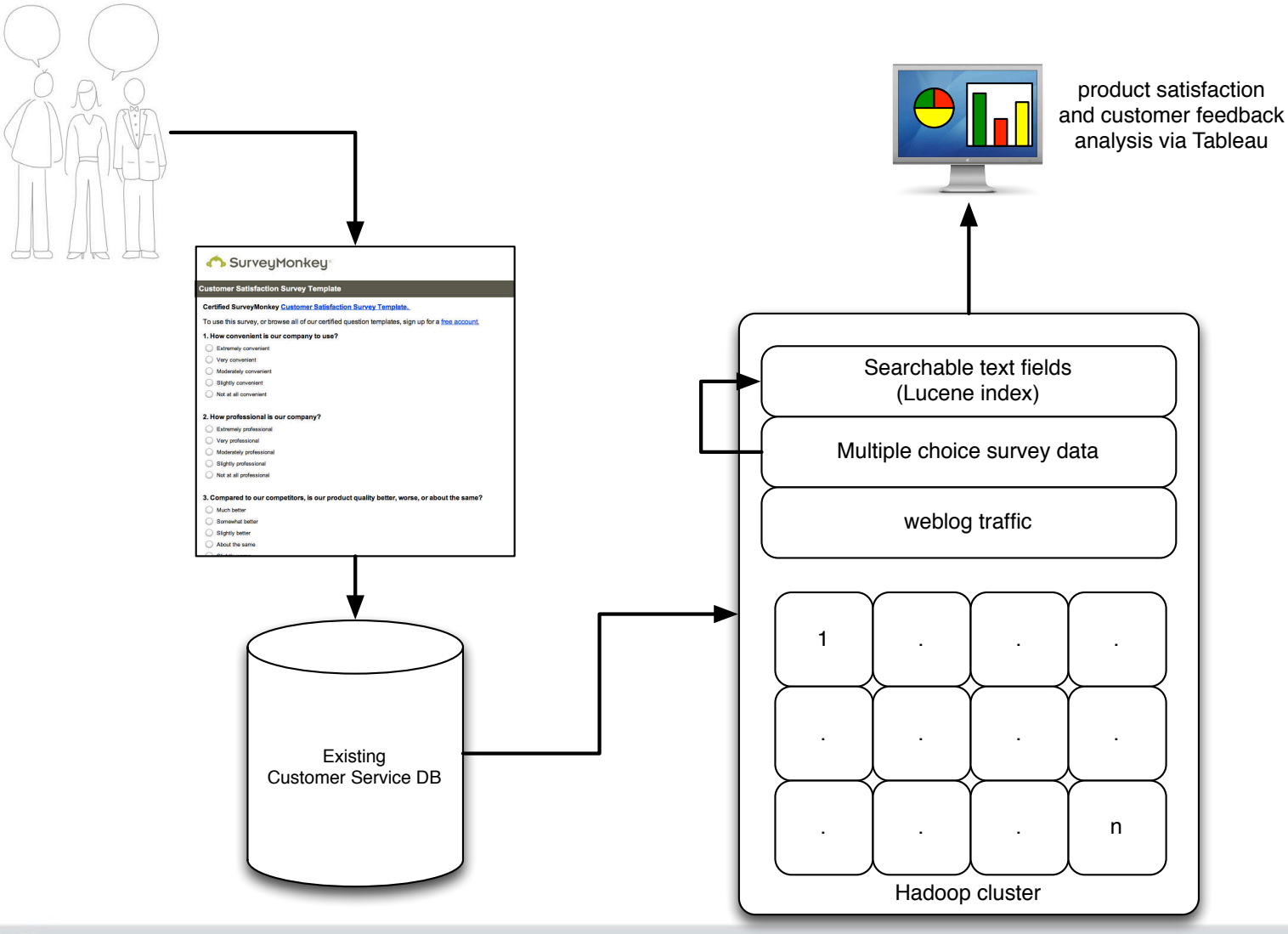
Process

- Parse the data into queryable format
- Explore & analyze using Hive, Pig, Mahout and other tools to discover value
- Label data and type information for compatibility and later discovery
- Pre-compute stats, groupings, patterns in data to accelerate analysis

Exchange

- Use visualization tools to facilitate exploration and find key insights
- Optionally move actionable insights into EDW or datamart

“Hardware Manufacturer” Unlocking Customer Survey Data



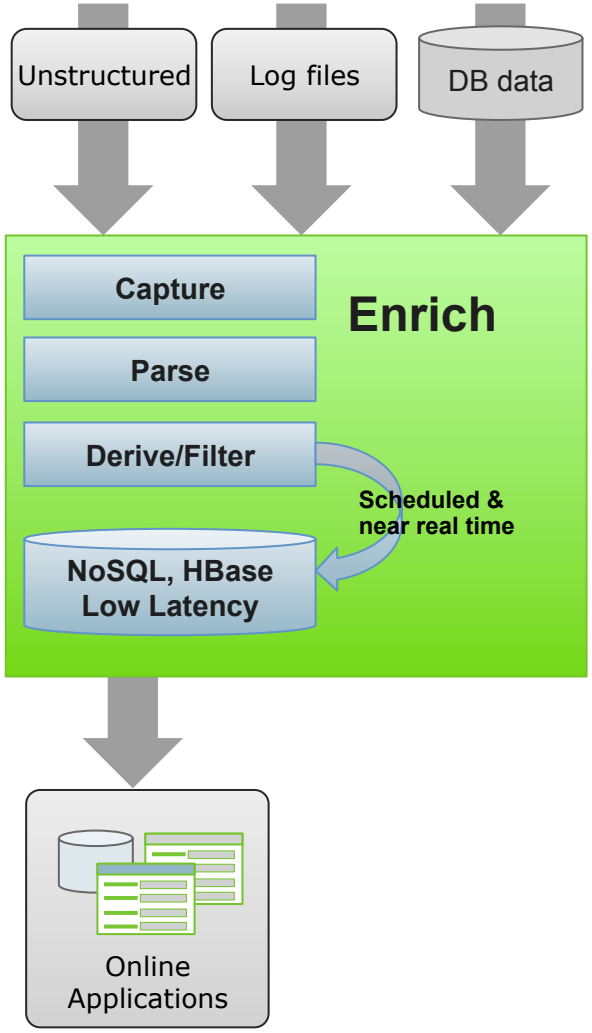
“Hardware Manufacturer” Key Benefits

- **Capture**
  - Store 10+ survey forms/year for > 3 years
  - Capture text, audio, and systems data in one platform
- **Process**
  - Unlock freeform text and audio data
  - Un-anonymize customers
  - Create HCatalog tables “customer”, “survey”, “freeform text”
- **Exchange**
  - Visualize natural satisfaction levels and groups
  - Tag customers as “happy” and report back to CRM database

Application Enrichment

Deliver Hadoop analysis to online apps

Refine Explore Enrich



Capture

- Capture data that was once too bulky and unmanageable

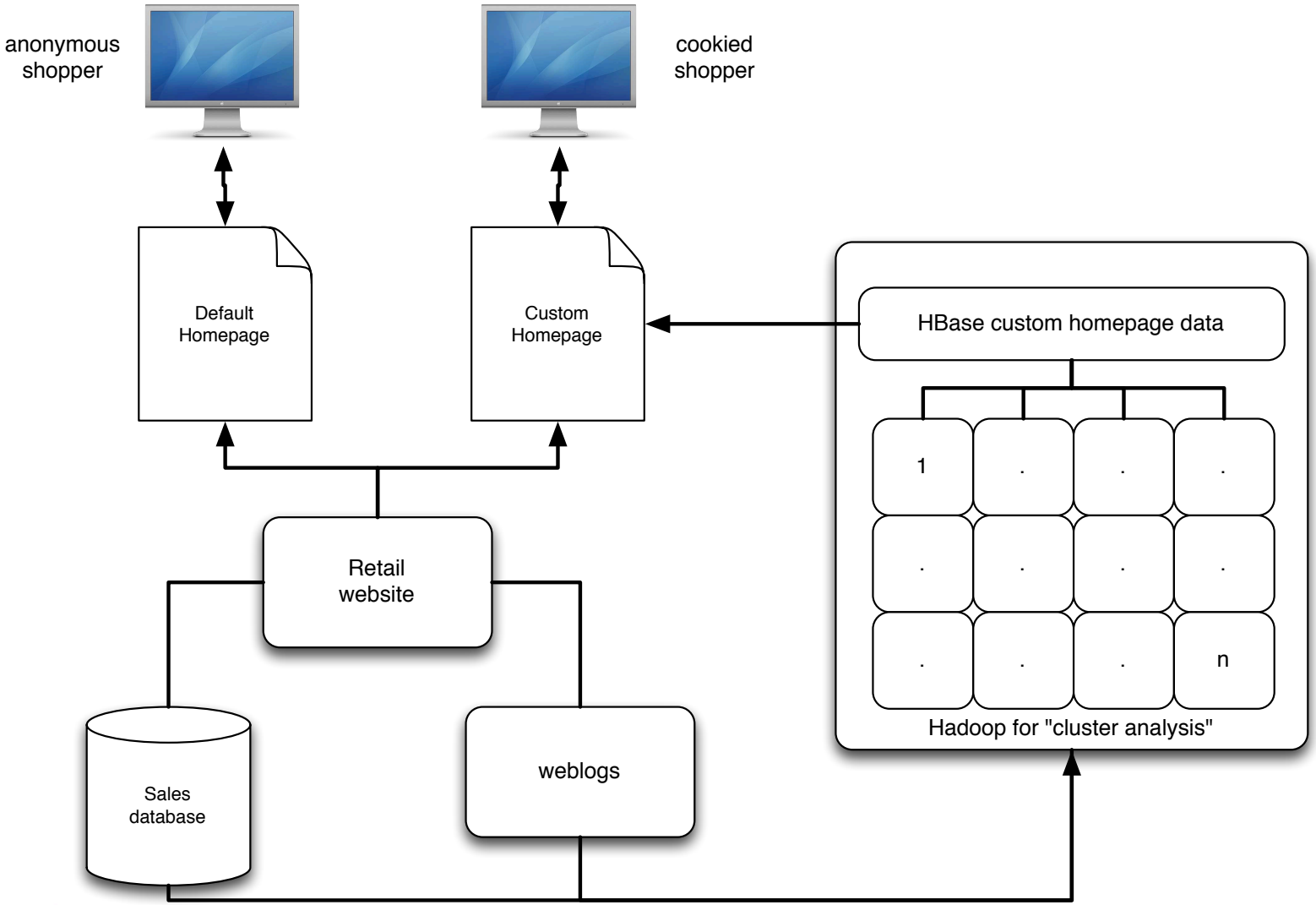
Process

- Uncover aggregate characteristics across data
- Use Hive Pig and Map Reduce to identify patterns
- Filter useful data from mass streams (Pig)
- Micro or macro batch oriented schedules

Exchange

- Push results to HBase or other NoSQL alternative for real time delivery
- Use patterns to deliver right content/offer to the right person at the right time

“Clothing Retailer” Custom Homepages



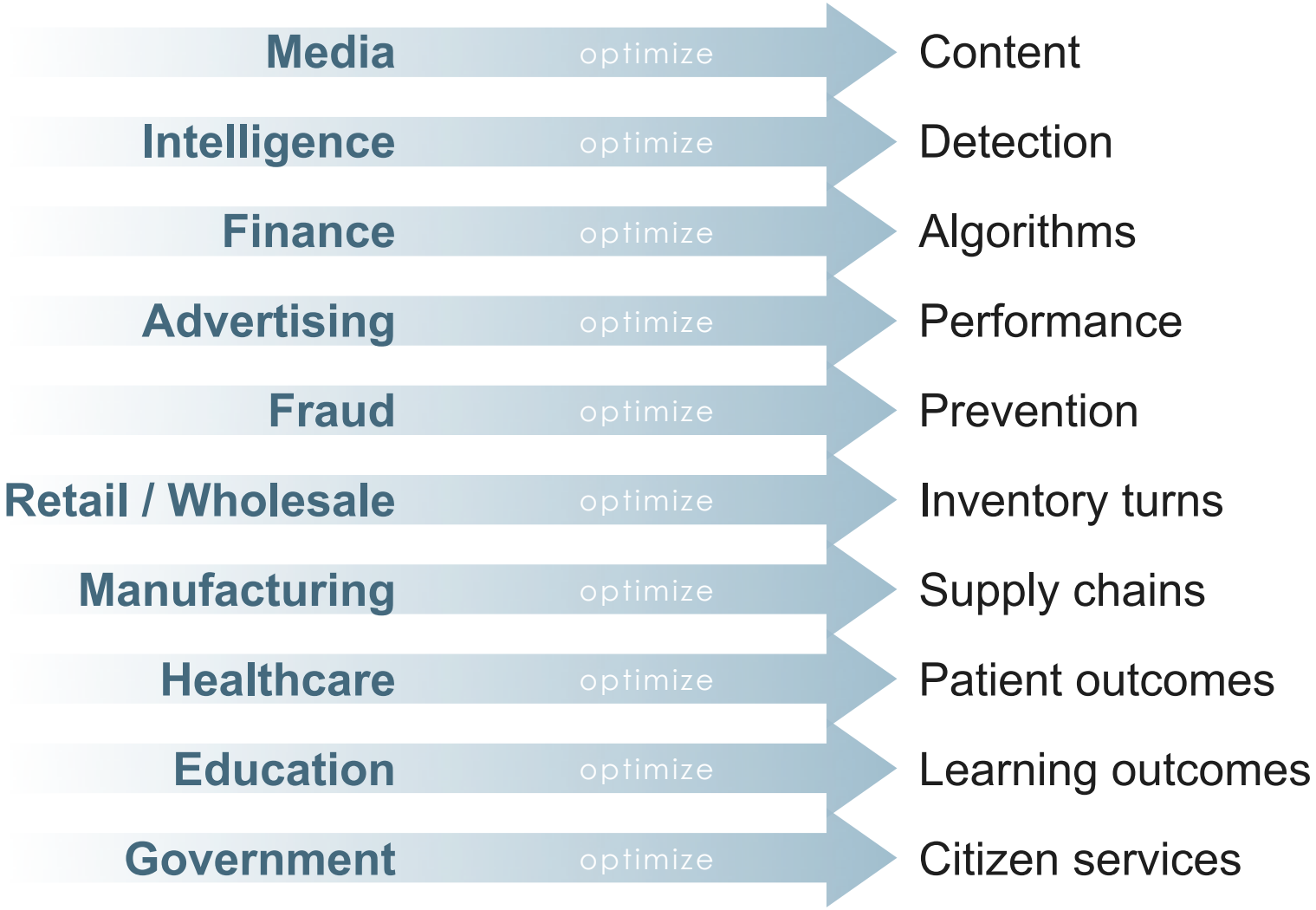
“Clothing Retailer” Key Benefits

- **Capture**
  - Capture web logs together with sales order history, customer master
- **Process**
  - Compute relationships between products over time
    - “people who buy shirts eventually need pants”
  - Score customer web behavior / sentiment
  - Connect product recommendations to customer sentiment
- **Exchange**
  - Load customer recommendations into HBase for rapid website service

Business Cases of Hadoop

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	• Smart Grid: Production Optimization	• Grid Failure Prevention • Smart Meters	• Individual Power Grid
Manufacturing	• Supply Chain Optimization	• Customer Churn Analysis	• Dynamic Delivery • Replacement parts
Healthcare & Payer	• Electronic Medical Records (EMPI)	• Clinical Trials Analysis	• Insurance Premium Determination

Goal: Optimize Outcomes at Scale



***Source**: Geoffrey Moore. Hadoop Summit 2012 keynote presentation.*

Customer: UC Irvine Medical Center

UCI Medical Center
University of California, Irvine

- UC Irvine Medical Center is ranked among the nation's best hospitals by U.S. News & World Report for the 12th year
- More than 400 specialty and primary care physicians
- Opened in 1976
- 422-bed medical facility

Optimizing patient outcomes while lowering costs

Current system, Epic holds 22 years of patient data, across admissions and clinical information

- Significant cost to maintain and run system
- Difficult to access, not-integrated into any systems, stand alone

Apache Hadoop sunsets legacy system and augments new electronic medical records

1. Migrate all legacy Epic data to Apache Hadoop
   - Replaced existing ETL and temporary databases with Hadoop resulting in faster more reliable transforms
   - Captures all legacy data not just a subset. Exposes this data to EMR and other applications
2. Eliminate maintenance of legacy system and database licenses
   - $500K in annual savings
3. Integrate data with EMR and clinical front-end
   - Better service with complete patient history provided to admissions and doctors
   - Enable improved research through complete information

Hortonworks

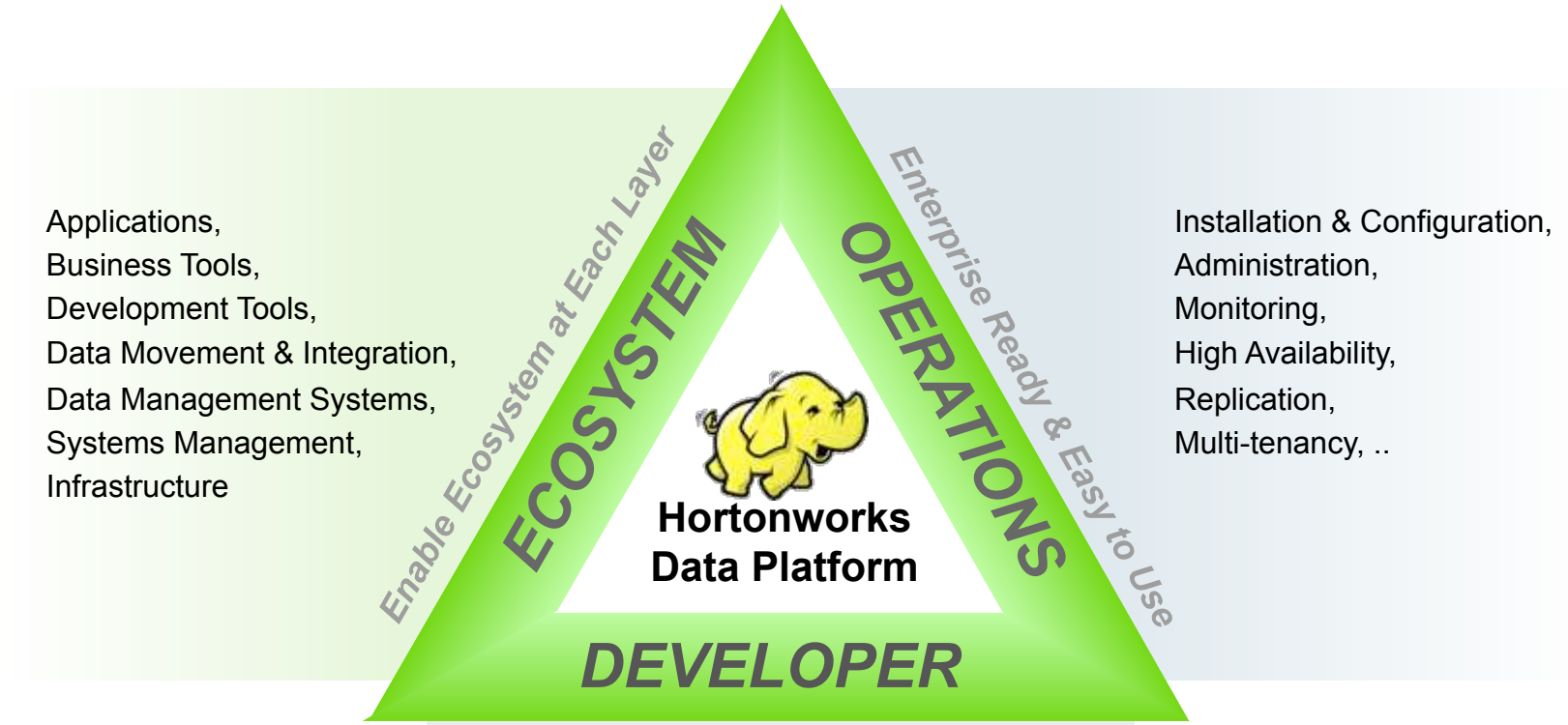
Hortonworks Data Platform

Hortonworks Vision & Role

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

1. **Be diligent stewards of the open source core**
2. **Be tireless innovators beyond the core**
3. **Provide robust data platform services & open APIs**
4. **Enable the ecosystem at each layer of the stack**
5. **Make the platform enterprise-ready & easy to use**

Enabling Hadoop as Enterprise Viable Platform



Delivering on the Promise of Apache Hadoop

Trusted

- Stewards of the core
- Original builders and operators of Hadoop
- 100+ years development experience
- Managed every viable Hadoop release
- HDP built on Hadoop 1.0

Open

- 100% open platform
- No POS holdback
- Open to the community
- Open to the ecosystem
- Closely aligned to Apache Hadoop core

Innovative

- Innovating current platform with HCatalog, Ambari, HA
- Innovating future platform with YARN, HA
- Complete vision for Hadoop-based platform

Hortonworks Apache Hadoop Expertise

Hortonworks represent more than 90 years combined Hadoop development experience

8 of top 15 highest lines of code* contributors to Hadoop of all time

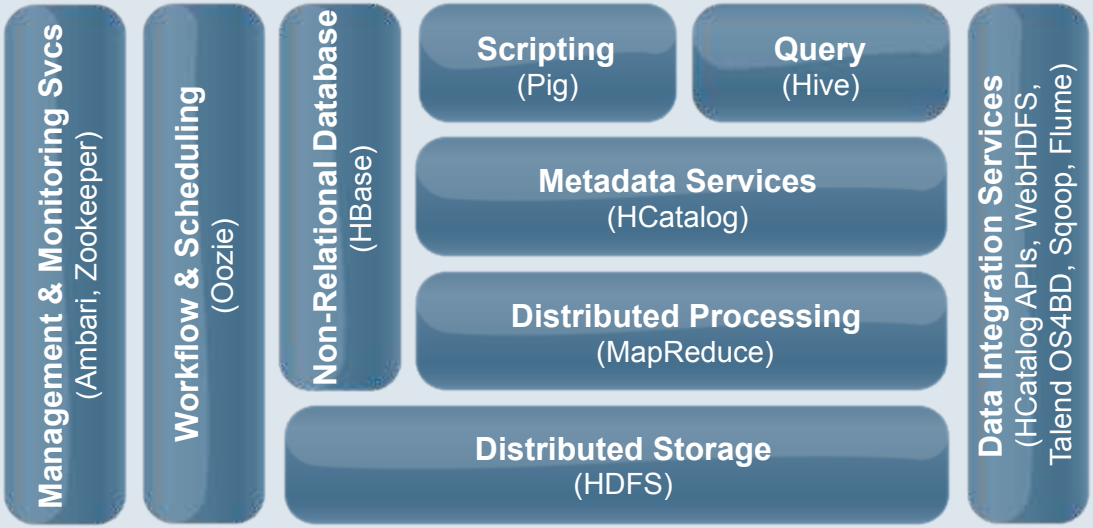
…and 3 of top 5

Hortonworkers are the builders, operators and core architects of Apache Hadoop

"Hortonworks' engineers spend all of their time improving open source Apache Hadoop. Other Hadoop engineers are going to spend the bulk of their time working on their proprietary software"

"We have noticed more activity over the last year from Hortonworks' engineers on building out Apache Hadoop's more innovative features. These include YARN, Ambari and HCatalog.."

Hortonworks Data Platform



Hortonworks Data Platform

Delivers enterprise grade functionality on a proven Apache Hadoop distribution to ease management, simplify use and ease integration into the enterprise

- **Simplify deployment** to get started quickly and easily
- Monitor, manage any size cluster with **familiar console** and tools
- Only platform to include **data integration services** to interact with any data source
- **Metadata services** opens the platform for integration with existing applications
- Dependable **high availability** architecture

The only 100% open source data platform for Apache Hadoop

Hortonworks

Demo

Why Hortonworks Data Platform?

ONLY Hortonworks Data Platform provides…

- **Tightly aligned to core Apache Hadoop development line**
  - Reduces risk for customers who may add custom coding or projects
- **Enterprise Integration**
  - HCatalog provides scalable, extensible integration point to Hadoop data
- **Most reliable Hadoop distribution**
  - Full stack high availability on v1 delivers the strongest SLA guarantees
- **Multi-tenant scheduling and resource management**
  - Capacity and fair scheduling optimizes cluster resources
- **Integration with operations, eases cluster management**
  - Ambari is the most open/complete operations platform for Hadoop clusters

Hortonworks Growing Partner Ecosystem

Hortonworks Support Subscriptions

Objective: help organizations to successfully develop and deploy solutions based upon Apache Hadoop

- **Full-lifecycle technical support available**
  - Developer support for design, development and POCs
  - Production support for staging and production environments
    - Up to 24x7 with 1-hour response times
- **Delivered by the Apache Hadoop experts**
  - Backed by development team that has released every major version of Apache Hadoop since 0.1
- **Forward-compatibility**
  - Hortonworks' leadership role helps ensure bug fixes and patches can be included in future versions of Hadoop projects

Hortonworks Training

Hortonworks UNIVERSITY

The expert source for Apache Hadoop training & certification

Role-based training for developers, administrators & data analysts

- Coursework built and maintained by the core Apache Hadoop development team.
- The “right” course, with the most extensive and realistic hands-on materials
- Provide an immersive experience into real-world Hadoop scenarios
- Public and Private courses available

Comprehensive Apache Hadoop Certification

- Become a trusted and valuable Apache Hadoop expert

Hortonworks CERTIFIED ADMINISTRATOR

Hortonworks CERTIFIED DEVELOPER

Next Steps?

1

Download Hortonworks Data Platform

hortonworks.com/download

2

Use the getting started guide

hortonworks.com/get-started

3

Learn more… get support

Hortonworks UNIVERSITY

- Expert role based training
- Course for admins, developers and operators
- Certification program
- Custom onsite options

hortonworks.com/training

Hortonworks Support

- Full lifecycle technical support across four service levels
- Delivered by Apache Hadoop Experts/Committers
- Forward-compatible

hortonworks.com/support

Thank You!

Questions & Answers

Follow: @hortonworks

Hortonworks

Q&A