

# Big Data: Making Sense of it all!

Jamie Engesser

E-mail : [jamie@hortonworks.com](mailto:jamie@hortonworks.com)



# Data Driven Business? Facts not Intuition!



*Data driven decisions are better decisions – its as simple as that. Using big data enables managers to decide on the basis of evidence rather than intuition. For that reason it has the potential to revolutionize management*

**Harvard Business Review  
October 2012**

# Web giants proved the ROI in **data products** applying data science to large amounts of data

Google  
AdWords

Prediction of click  
through rates

Customers Who Bought Items in Your Shopping Cart Also Bought



Amazon: 35% of  
product sales come  
from product  
recommendations



Netflix: 75% of streaming  
video results from  
recommendations



**CENTERS FOR DISEASE<sup>TM</sup>  
CONTROL AND PREVENTION**



+28% in last 24 hours, -3% in last 4 hours, -49% in last 2 hours | **Cholera** : +75% in last 24 hours, -12% in last 4 hours, +37% in last 2 hours | **Common Cold**

# Welcome!

We are tracking disease trends, 140 characters at a time

The numbers below represent data for the past two weeks. Full historical data is being maintained but is not publicly available at this time.

**4,620,536**

tweets gathered from Twitter's Streaming API. All of them match at least one of the 234 condition terms currently tracked across 27 conditions.

**76,155 (1%)**

tweets with a sensor-based location [\(read more about how we calculate this\)](#)

**3,177,088 (68%)**

tweets with a popular user profile location [\(read more about how we calculate this\)](#)

## Conditions by Tweet Count

std acute respiratory  
illness influenza pertussis  
gastroenteritis common cold  
tuberculosis rabies pneumonia  
tetanus malaria varicella polio  
dengue meningitis tick borne  
disease anthrax cholera  
mumps measles mosquito  
borne disease typhoid  
smallpox legionnaires disease  
enterovirus diphtheria yellow  
fever

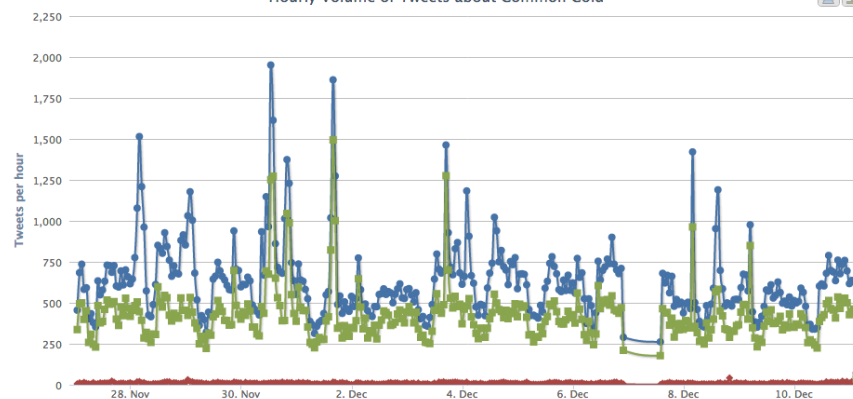
## Top 20 Tweet Locations

orlando, fl johor bahru,  
johore los angeles, ca  
georgia, us chicago, il south  
carolina, us manhattan, ny  
ohio, us texas  
new york, ny  
houston, tx  
bandar kuala  
lumpur toro  
boston, ma ala  
carolina, us  
lon

## Top 20 User Locations

indonesia london  
philippines singapore new  
york canada jakarta online  
usa uk malaysia england

Hourly Volume of Tweets about Common Cold



☒ Common Cold
 ☒ Common Cold tweets w/ tweet location
 ☒ Common Cold tweets w/ no urls and no hashtags

☒ Graph volume and geotweets only (default)
 ☐ Graph condition terms too
 ☐ Graph qualifier matches too
 ☐ Graph everything

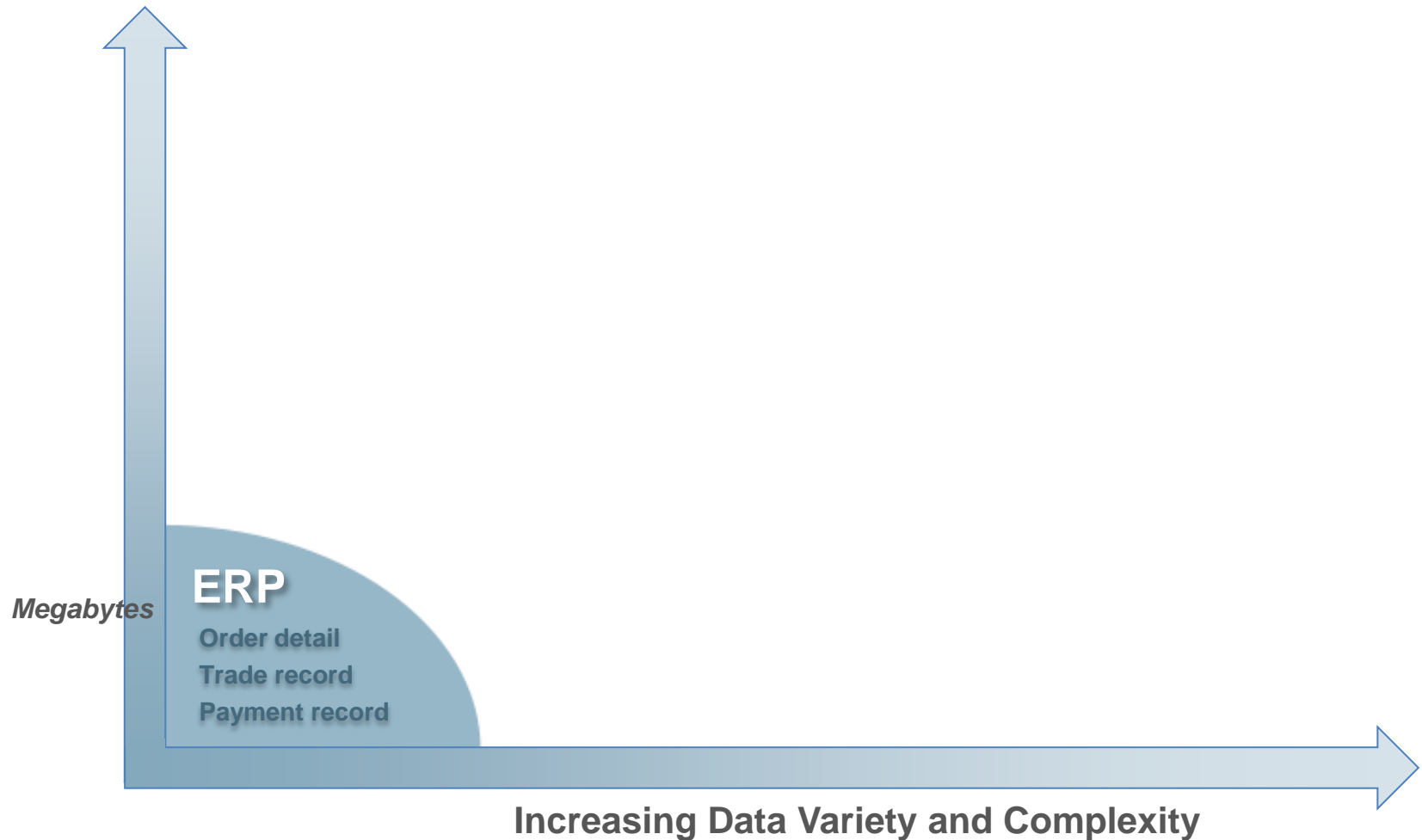
Highcharts.com

# Big Data: Optimize Outcomes at Scale

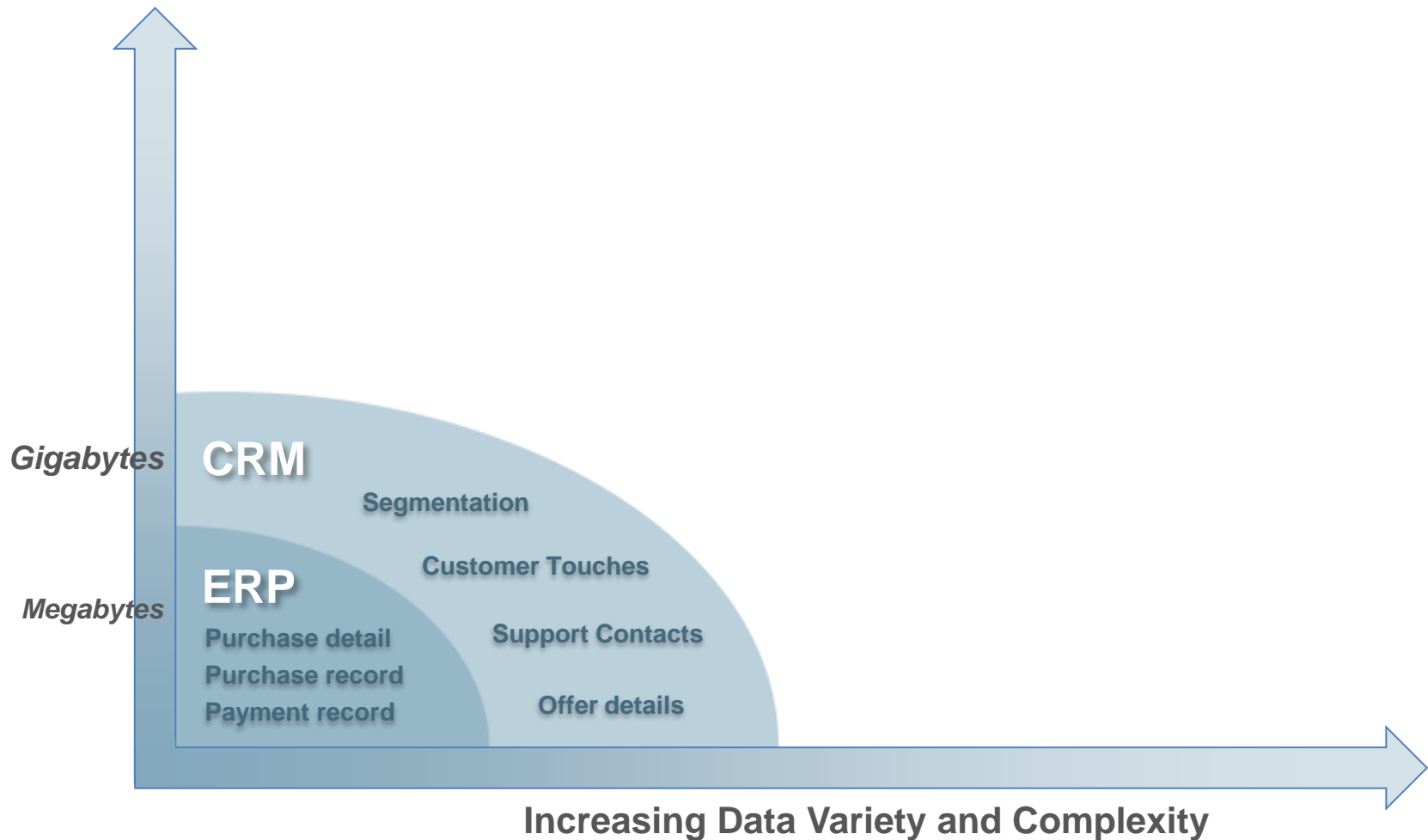


**Source:** Geoffrey Moore. *Hadoop Summit 2012 keynote presentation.*

# Analytics started with basic transaction history...

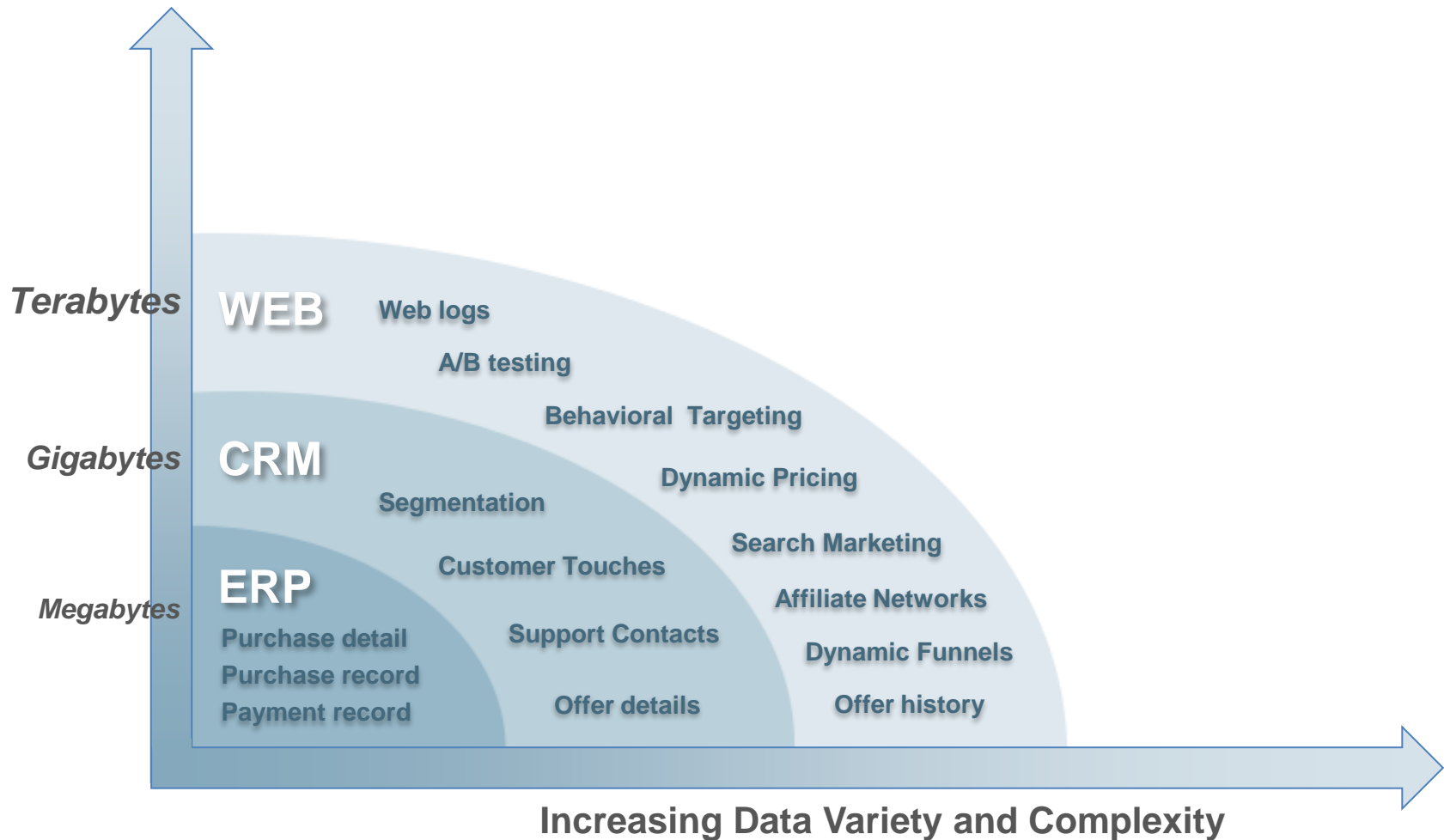


# then we added customer information...

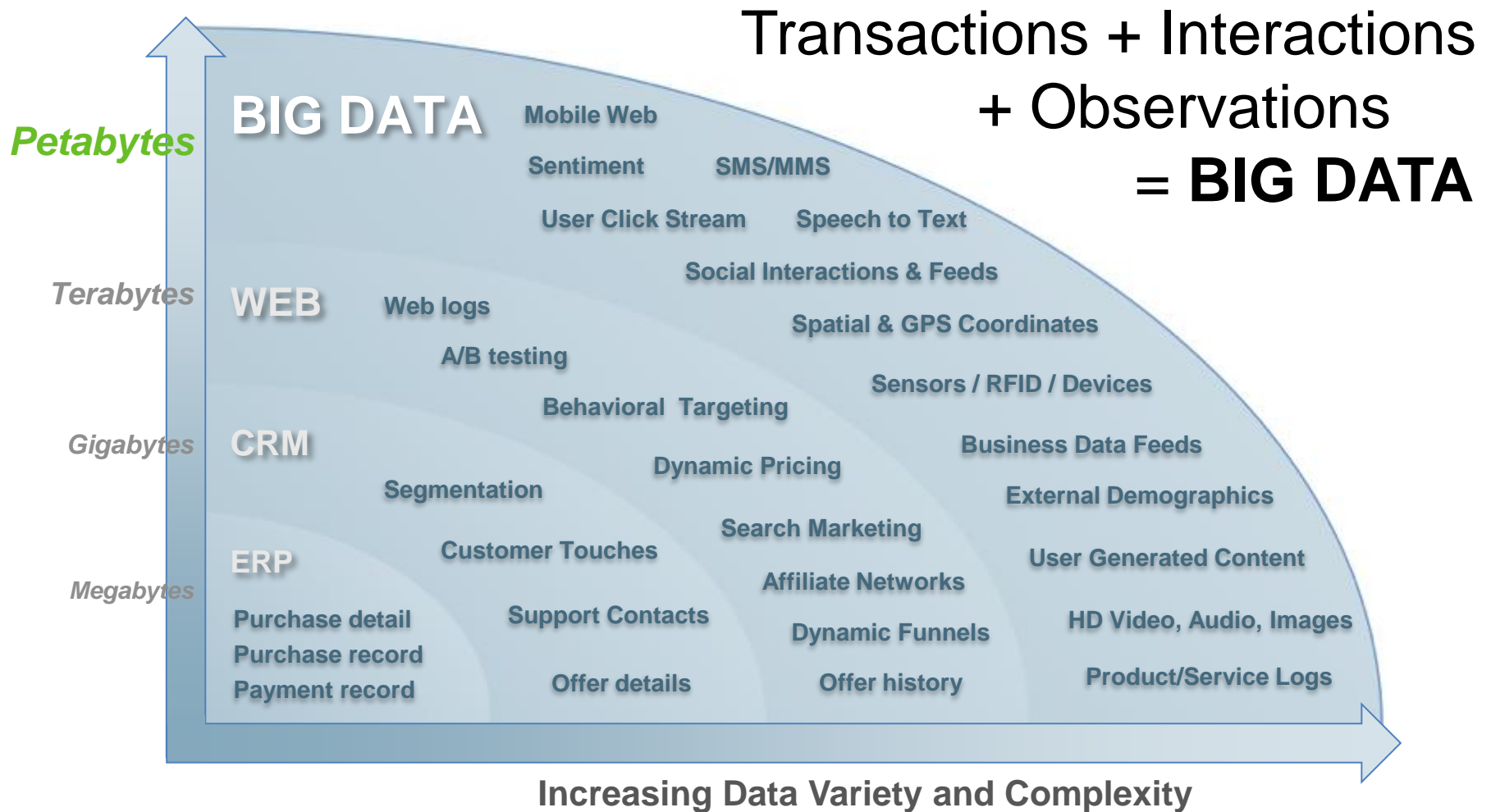





# and the web started to impact...




# Big Data: Organizational Game Changer





# A little history... it's 2005



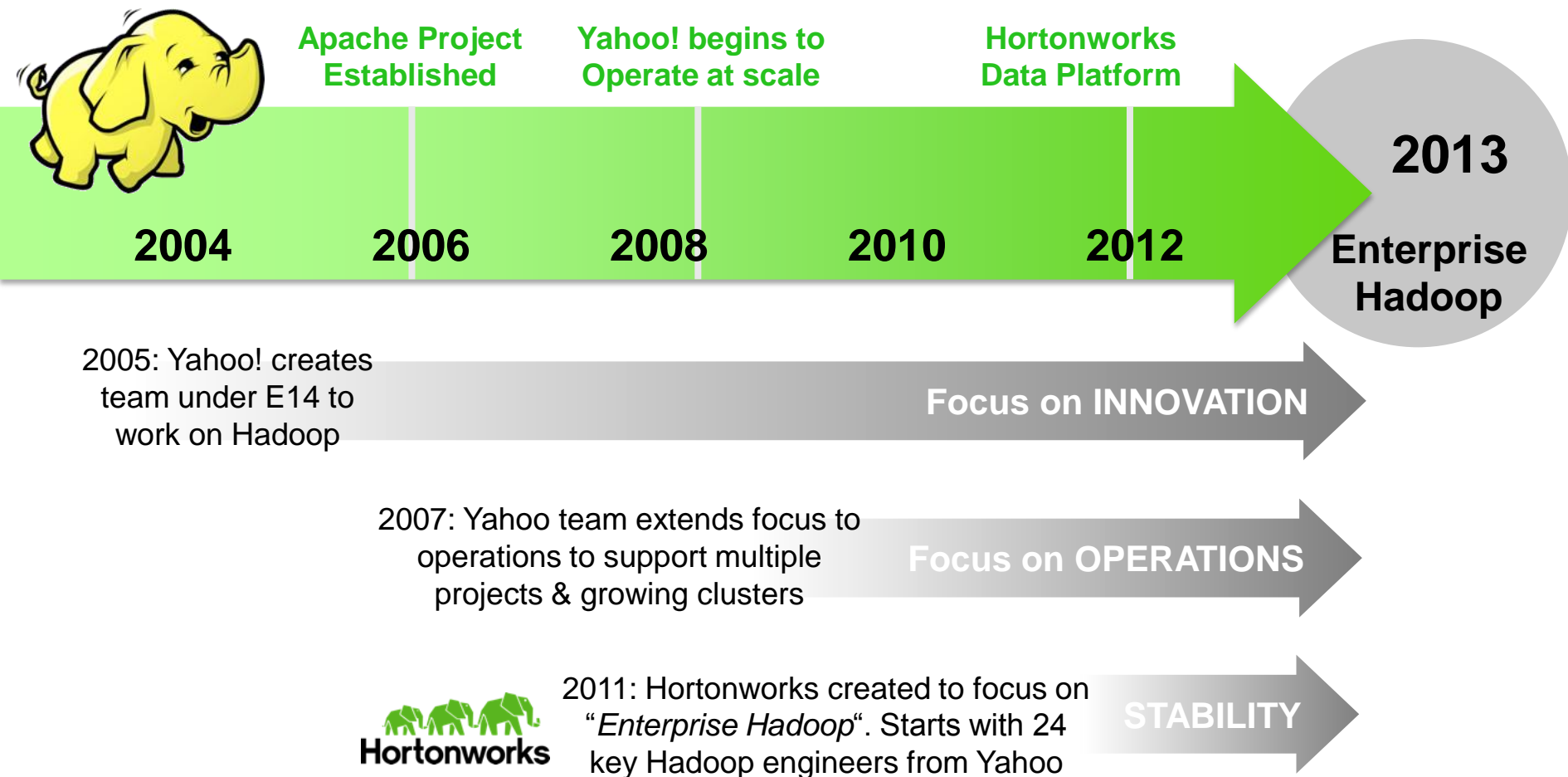


24	AGE	26
147	WEIGHT	147
HEIGHT		
5 ft.6½ in.		5 ft.7 ½ in.
REACH		
73 in.		72 in.
CHEST NORMAL		
36 ½ in.		41 in.
CHEST EXPANDED		
38 ½ in.		43 in.
WAIST		
29 ½ in.		28 in.
NECK		
16 in.		16 in.
FIST		
11 in.		11 ½ in.
CALF		
15 ½ in.		14 in.
BICEPS		
13 ½ in.		16 ½ in.





# A Brief History of Apache Hadoop



# Apache Hadoop: Center of Big Data Strategy



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

Storage

## HDFS



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

Processing

## Map Reduce



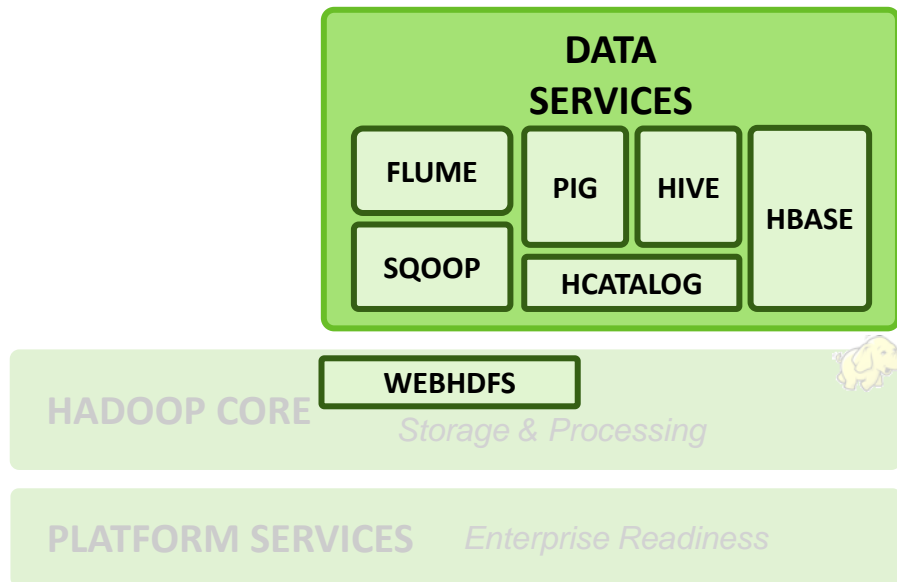
- Splits a task across processors “near” the data & assembles results
- Self-Healing, High Bandwidth Clustered Storage

## 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

# Hadoop is more?

# Data Services for Full Data Lifecycle



Provide **data services** to store, process & access data in many ways

## Unique Focus Areas:

- **Apache HCatalog**  
Metadata services for consistent table access to Hadoop data
- **Apache Hive**  
Explore & process Hadoop data via SQL & ODBC-compliant BI tools
- **Apache HBase**  
NoSQL database for Hadoop
- **WebHDFS**  
Access Hadoop files via scalable REST API
- **Talend Open Studio for Big Data**  
Graphical data integration tools

# Metadata Service & Table-level Abstractions

## Apache HCatalog provides flexible metadata services across tools and external access

- **Consistency** of metadata and data models across tools (MapReduce, Pig, HBase and Hive)
- **Accessibility**: share data as tables in and out of HDFS
- **Availability**: enables flexible, thin-client access via REST API

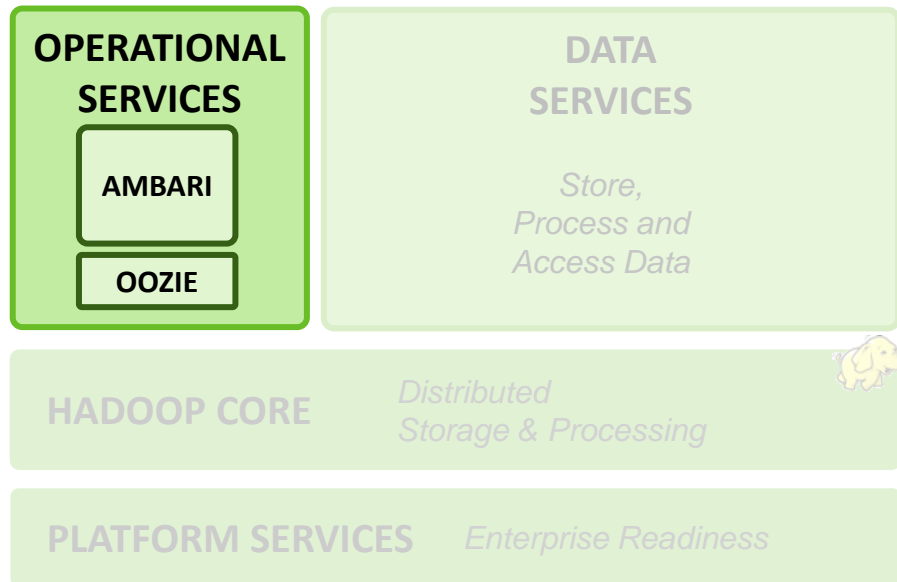


*Shared table and schema management opens the platform*

- |                         |   |                  |
|-------------------------|---|------------------|
| • Raw Hadoop data       | → | Table access     |
| • Inconsistent, unknown | → | Aligned metadata |
| • Tool specific access  | → | REST API         |



# Operational Services for Ease of Use

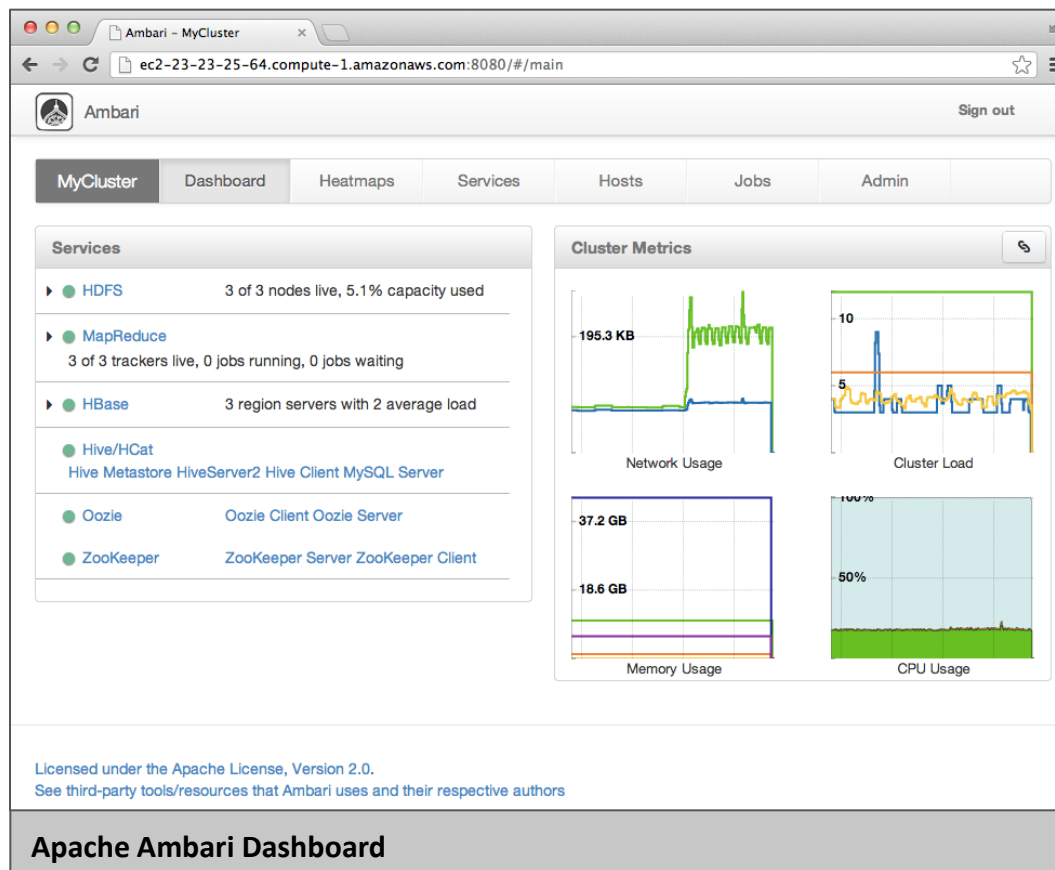


Include complete ***operational services*** for productive operations & management

## Unique Focus Area:

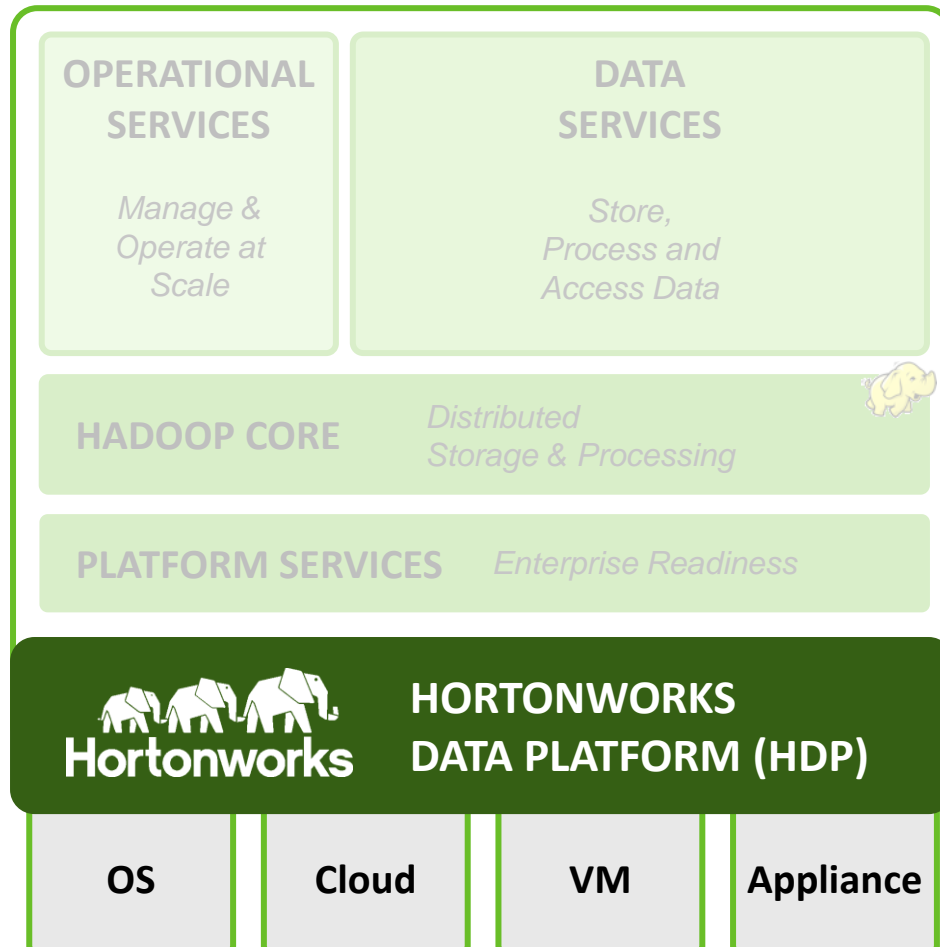
- **Apache Ambari:**  
Provision, manage & monitor a cluster; complete REST APIs to integrate with existing operational tools; job & task visualizer to diagnose issues

# New Ambari Features



- **Job Diagnostics**  
Visualize and troubleshoot Hadoop job execution and performance
- **Cluster History**  
View historical job execution & performance
- **Instant Insight**  
View health of Core Hadoop (HDFS, MapReduce) and related projects
- **Cluster Navigation**  
“Quick link” buttons jump into namenode web UI for a server
- **REST interface**  
provides external access to Ambari for existing tools. Facilitates integration with Microsoft System Center and Teradata Viewpoint

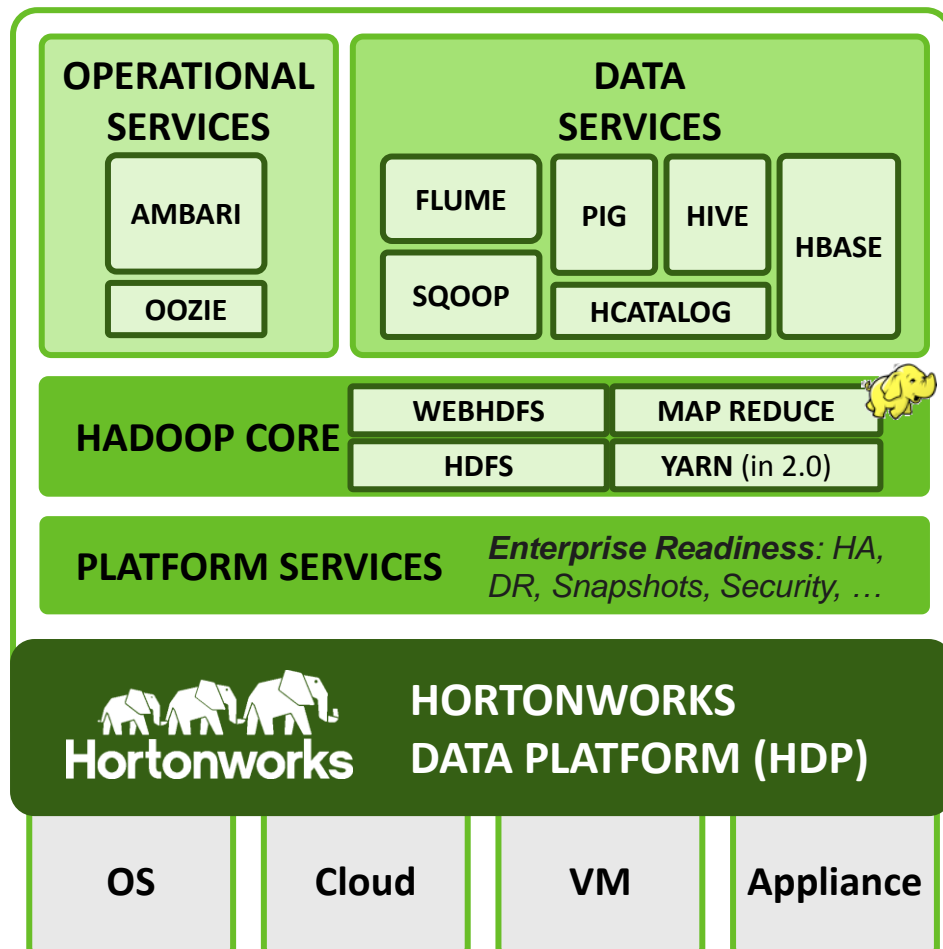
# Deployable Across a Range of Options



Hadoop allows you to ***deploy seamlessly*** across any deployment option

- Linux & *Windows*
- Azure, Rackspace & other clouds
- Virtual platforms
- Big data appliances

# Enterprise Hadoop Distribution

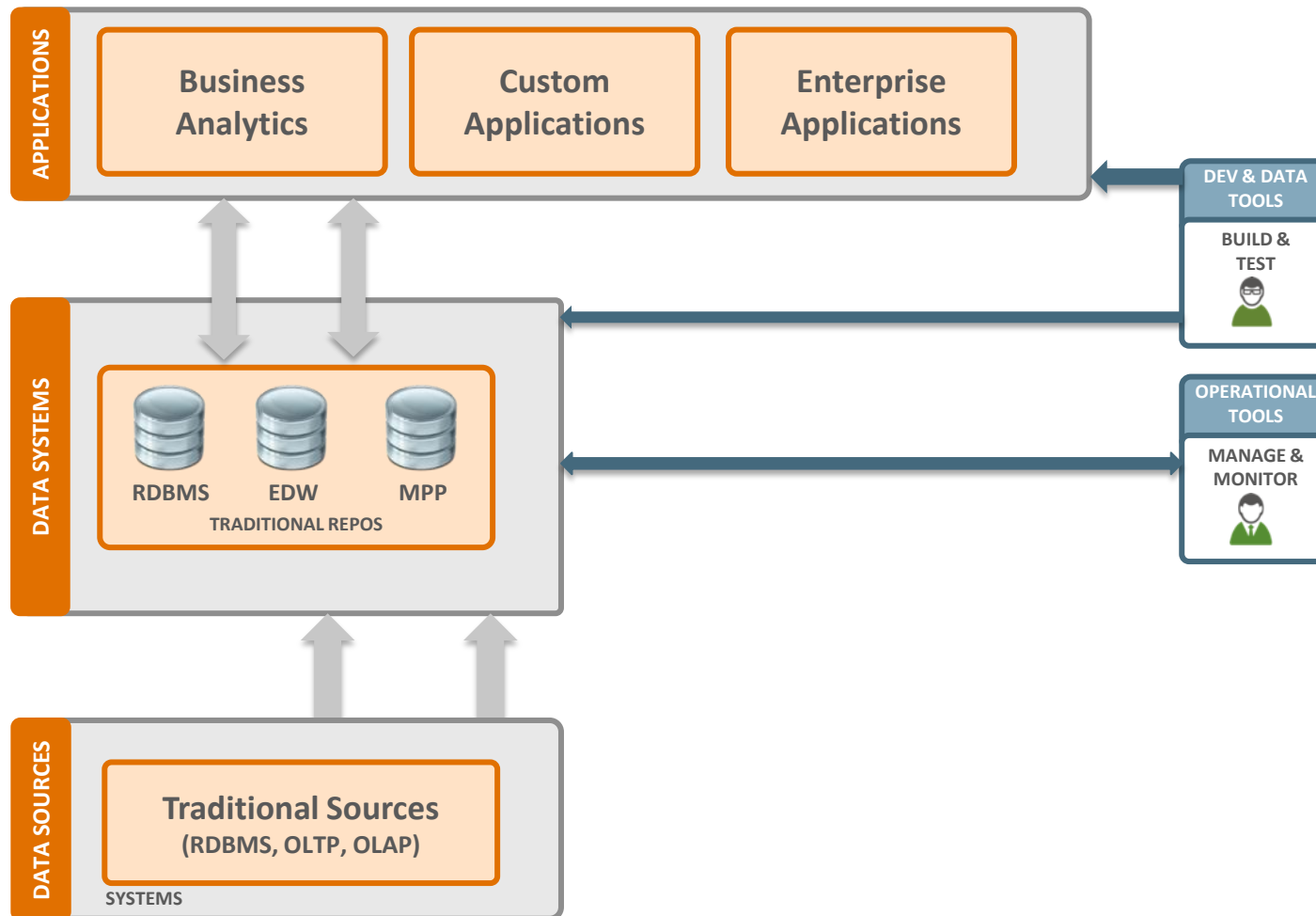


## Hortonworks Data Platform (HDP) *Enterprise Hadoop*

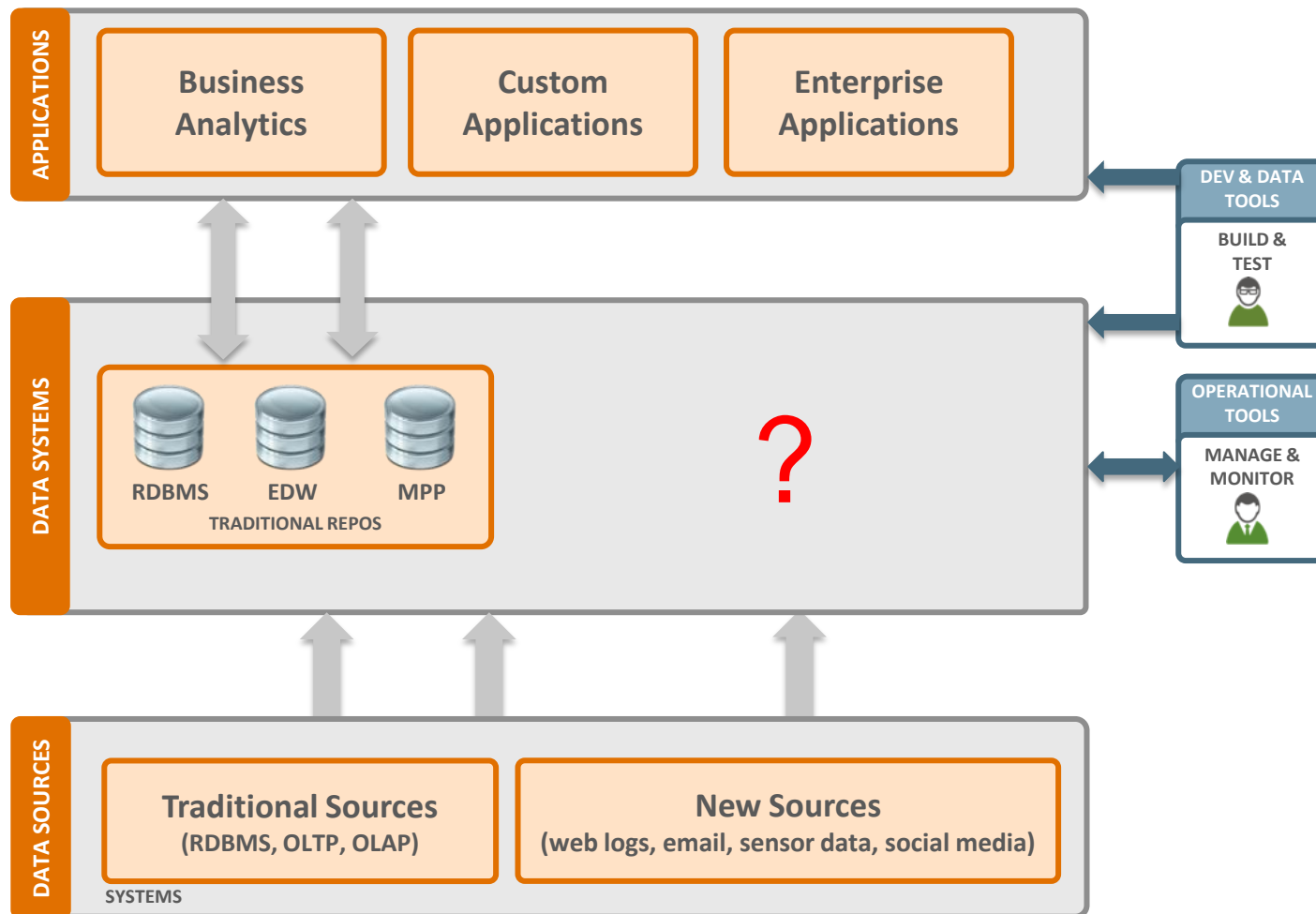
- The **ONLY** 100% open source and complete distribution
- Enterprise grade, proven and tested at scale
- Ecosystem endorsed to ensure interoperability

# Where does it fit in the enterprise?

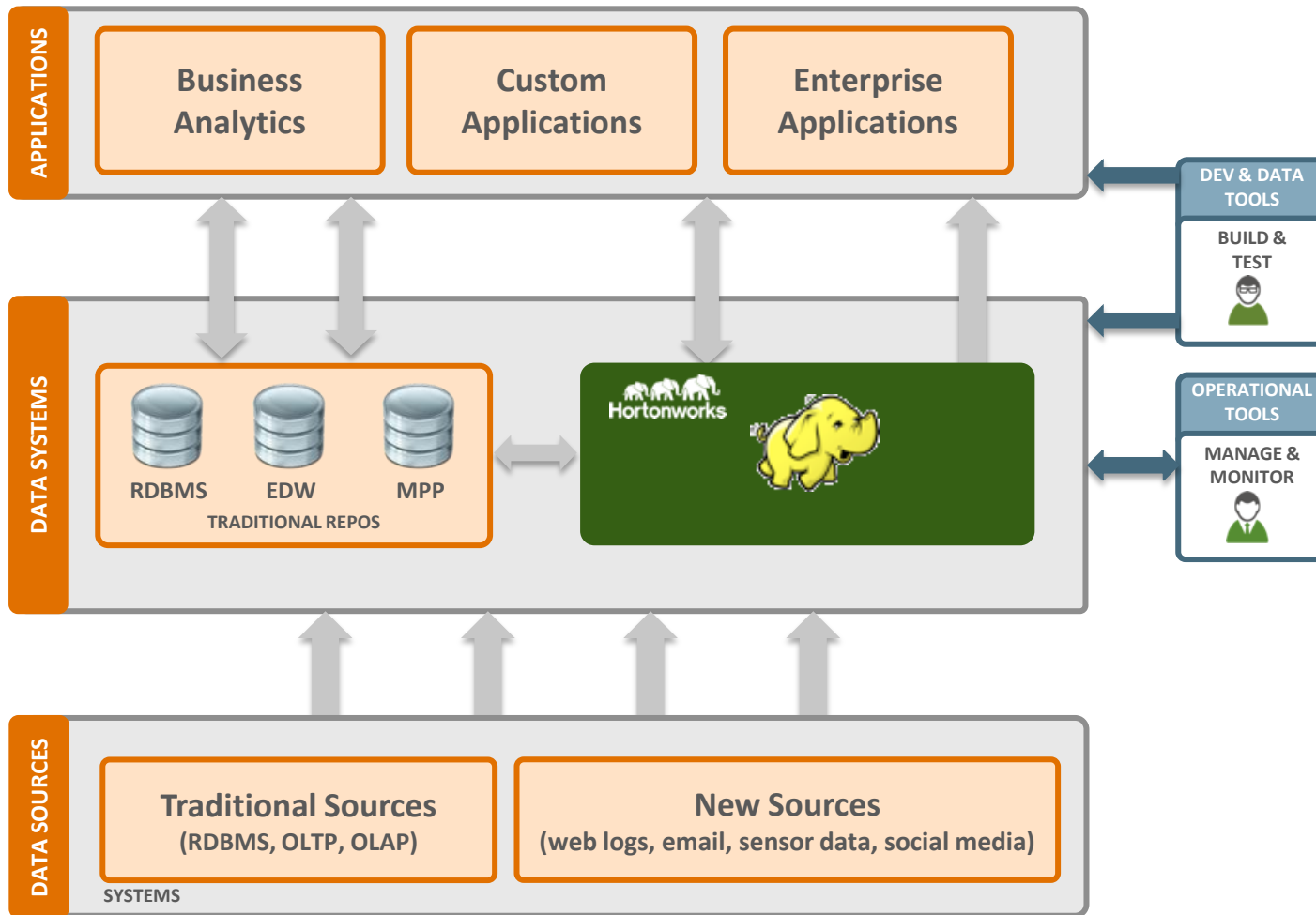
# Existing Data Architecture



# Existing Data Architecture

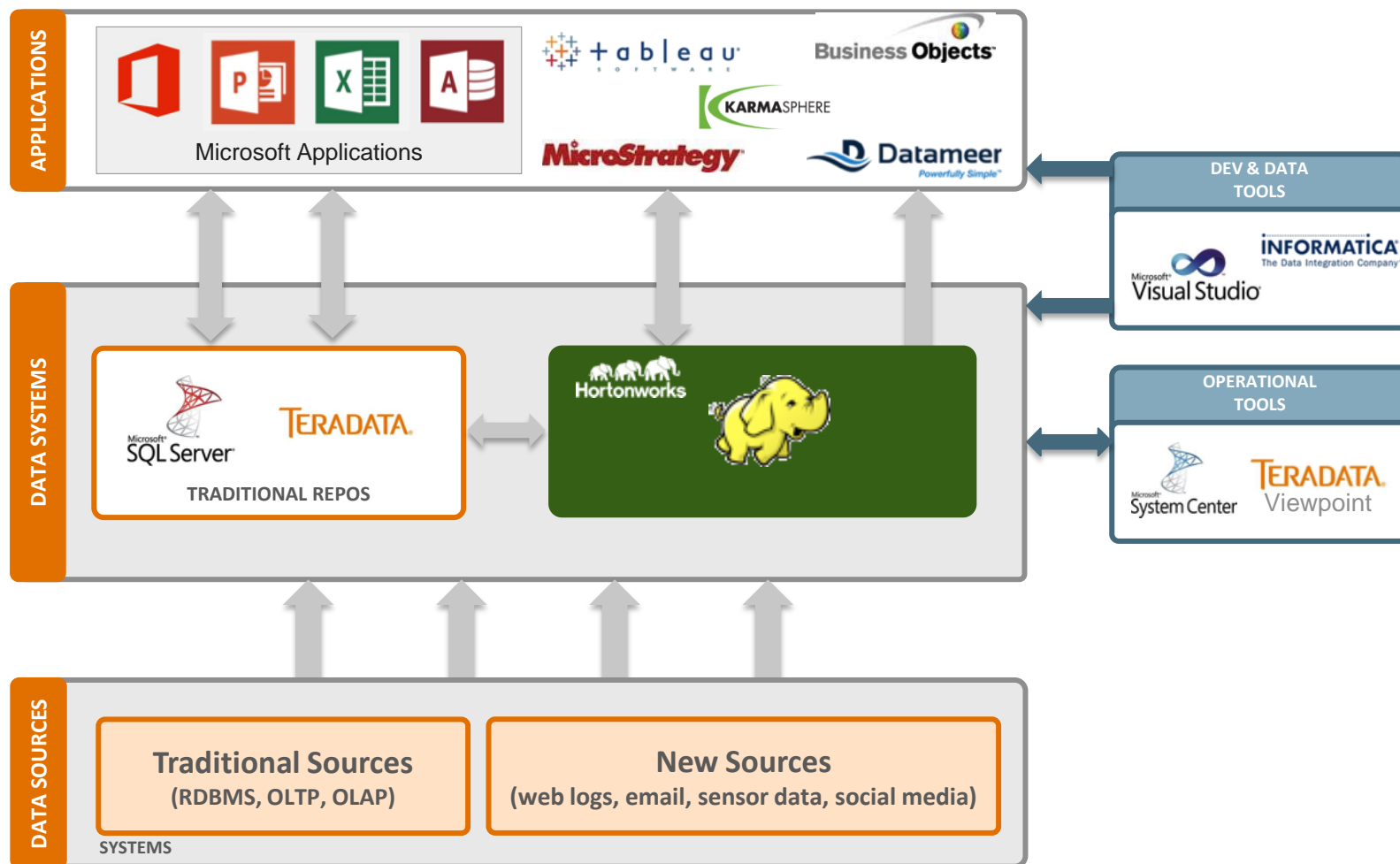


# An Emerging Data Architecture





# Interoperating With Your Tools



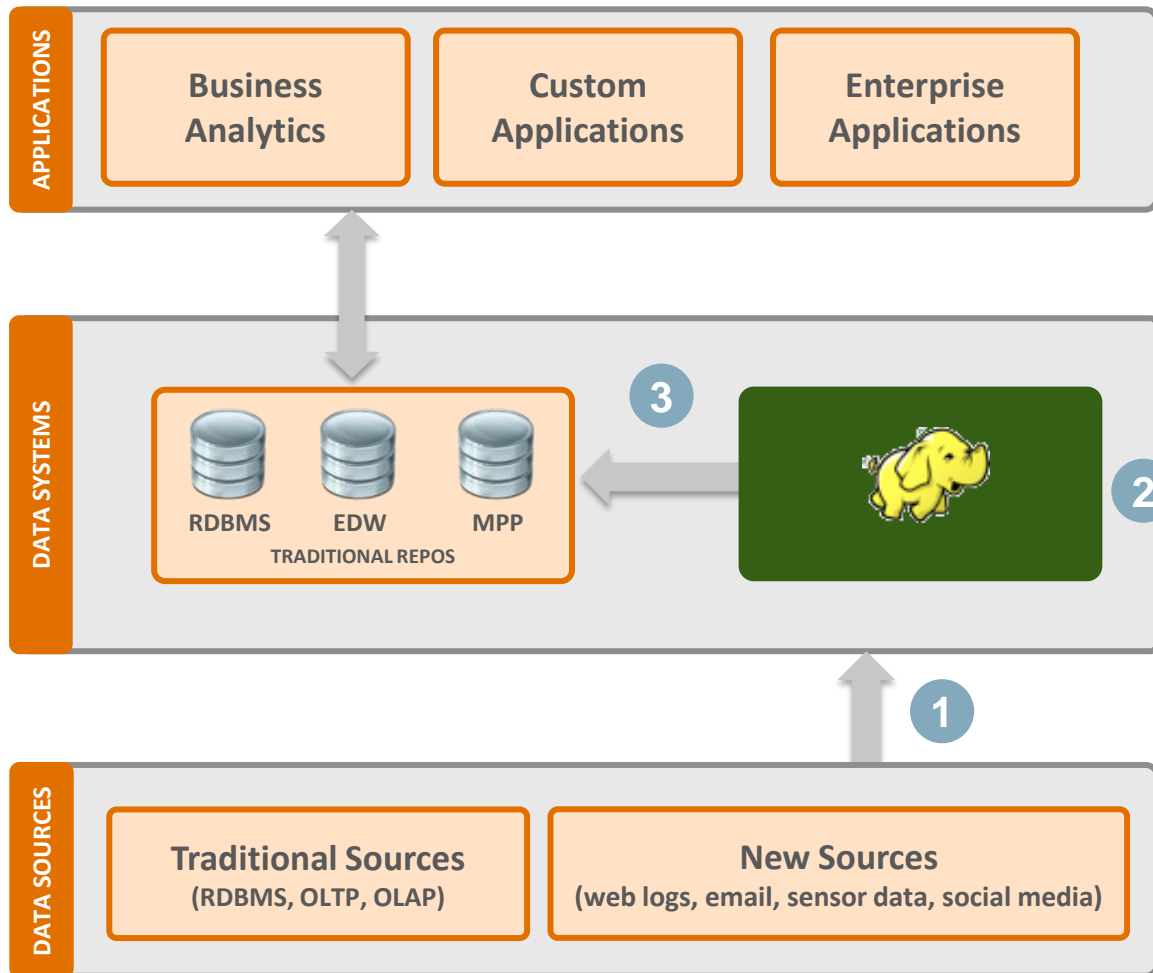
# Data Refinement

# Operational Data Refinery

Refine

Explore

Enrich



Collect data and apply a known algorithm to it in trusted operational process

## 1 Capture

Capture all data – New and Traditional

## 2

## Process

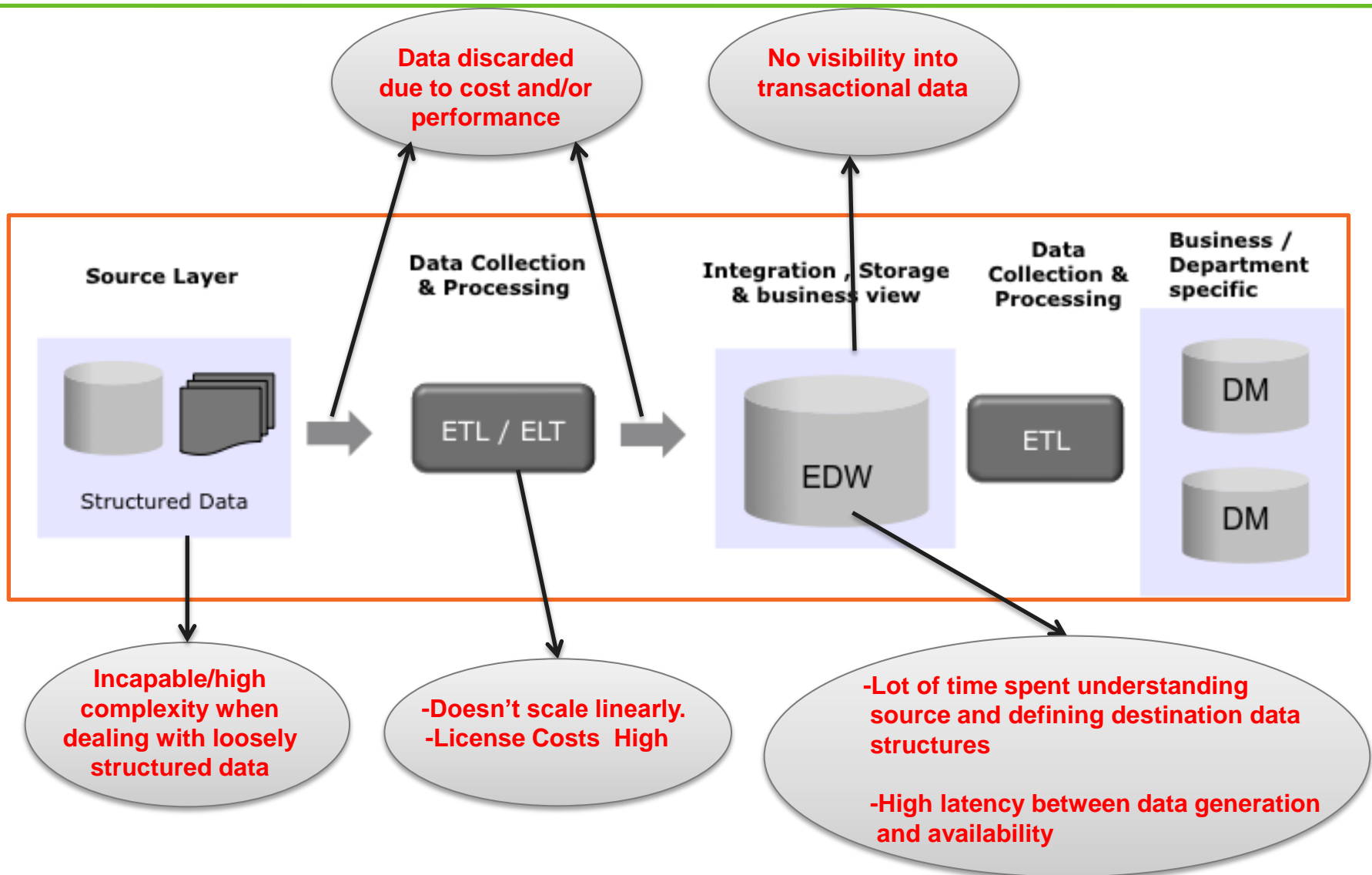
Parse, cleanse, apply structure & transform

## 3

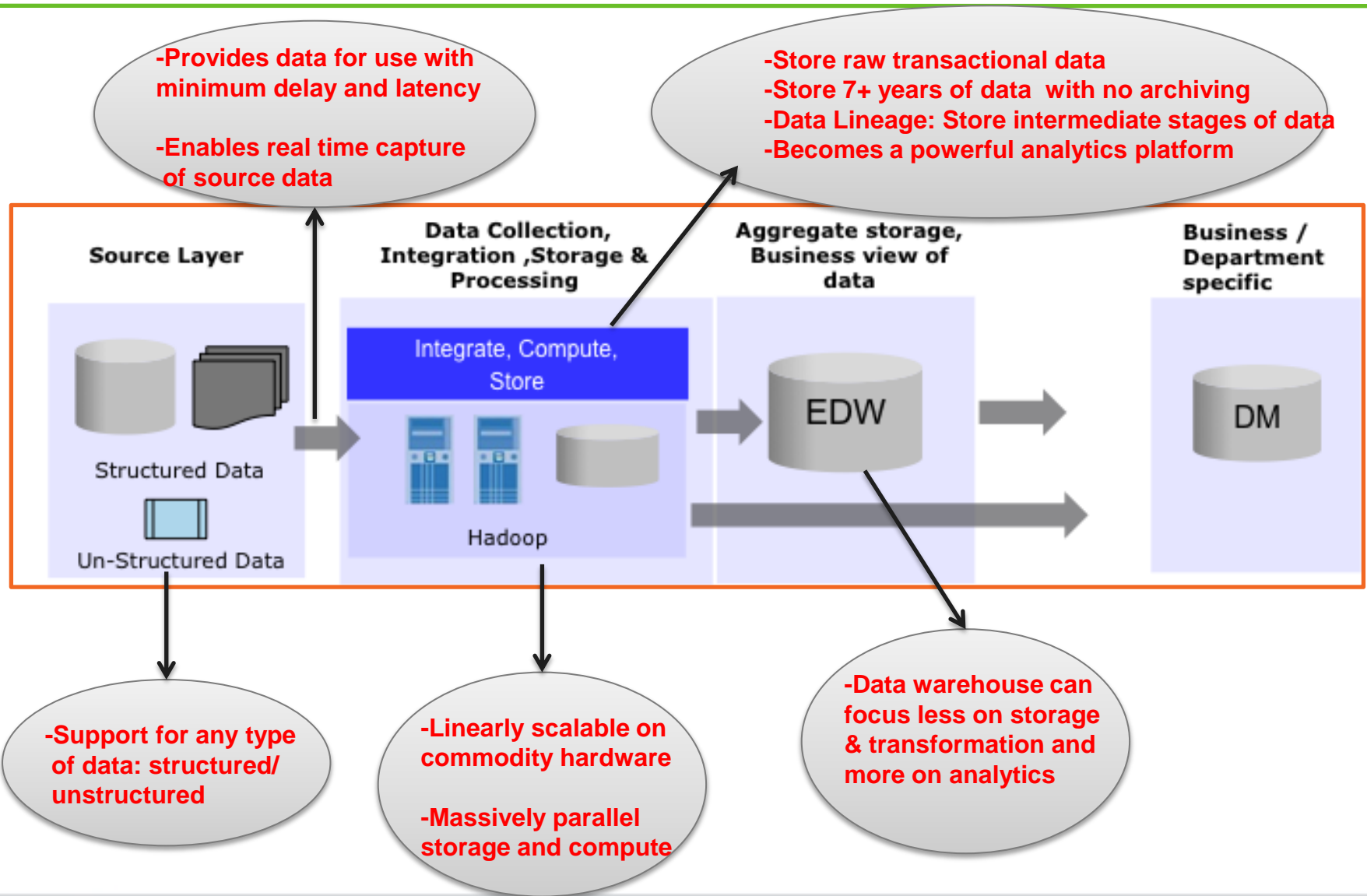
## Exchange

Push to existing data warehouse for use with existing analytic tools

# Challenges with a Traditional ETL Platform

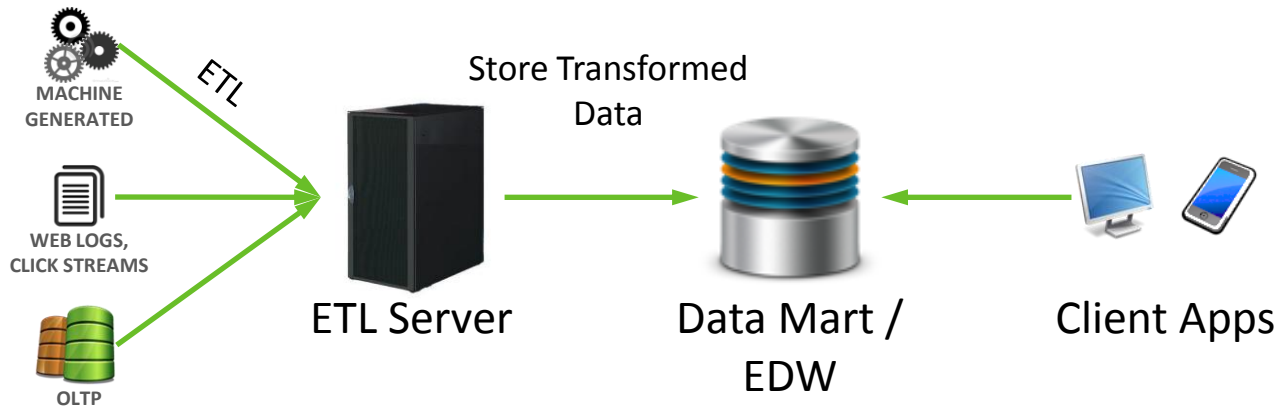


# HDP Based ETL Platform

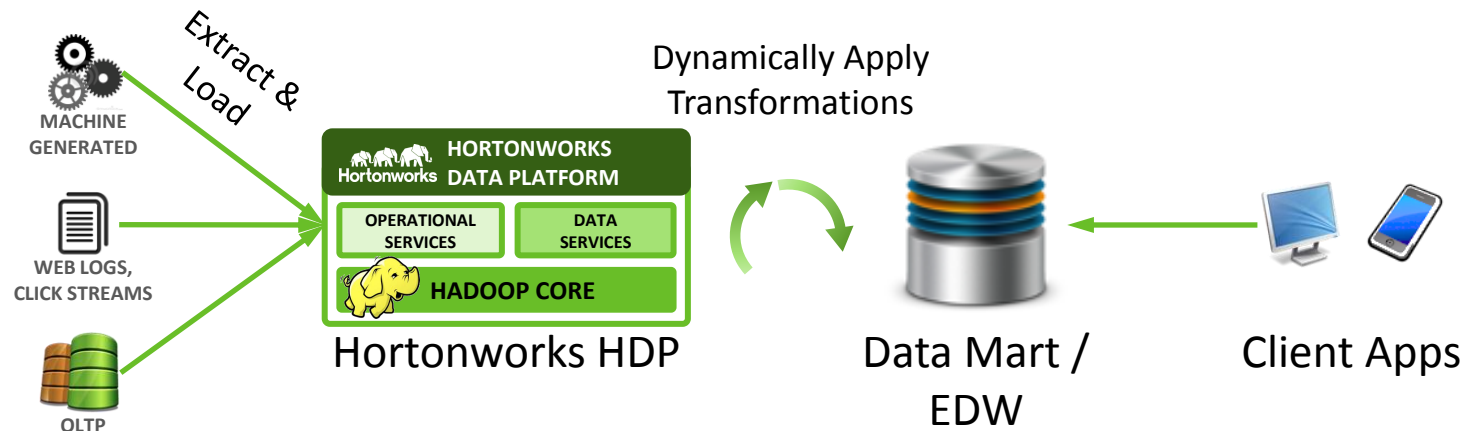


# Key Capability in Hadoop: Late binding

With traditional ETL, structure must be agreed upon far in advance and is difficult to change.



With Hadoop, capture all data, structure data as business need evolve.



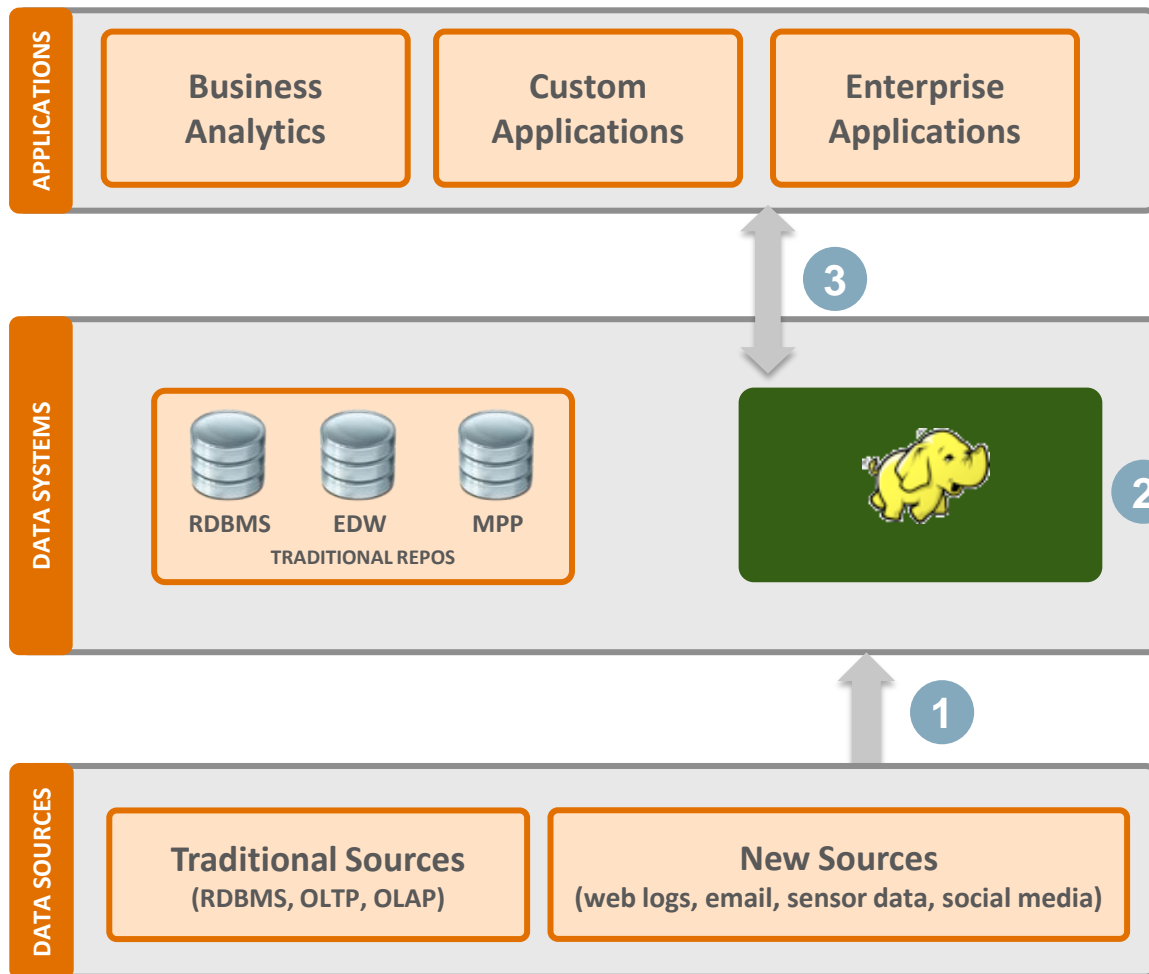
# Data Exploration

# Big Data Exploration & Visualization

Refine

Explore

Enrich



Collect data and perform iterative investigation for value

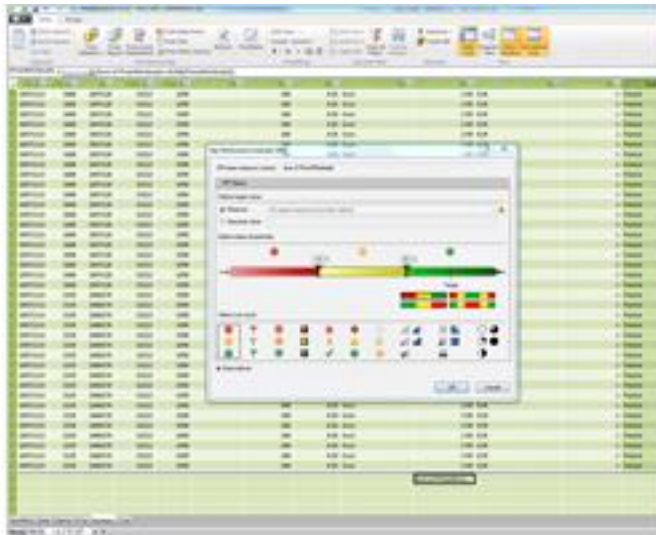
- 1 Capture**  
Capture all data
- 2 Process**  
Parse, cleanse, apply structure & transform
- 3 Exchange**  
Explore and visualize with analytics tools supporting Hadoop



# Visualization Tooling

- **Robust visualization and business tooling**
- **Ensures scalability when working with large datasets**

Native Excel support



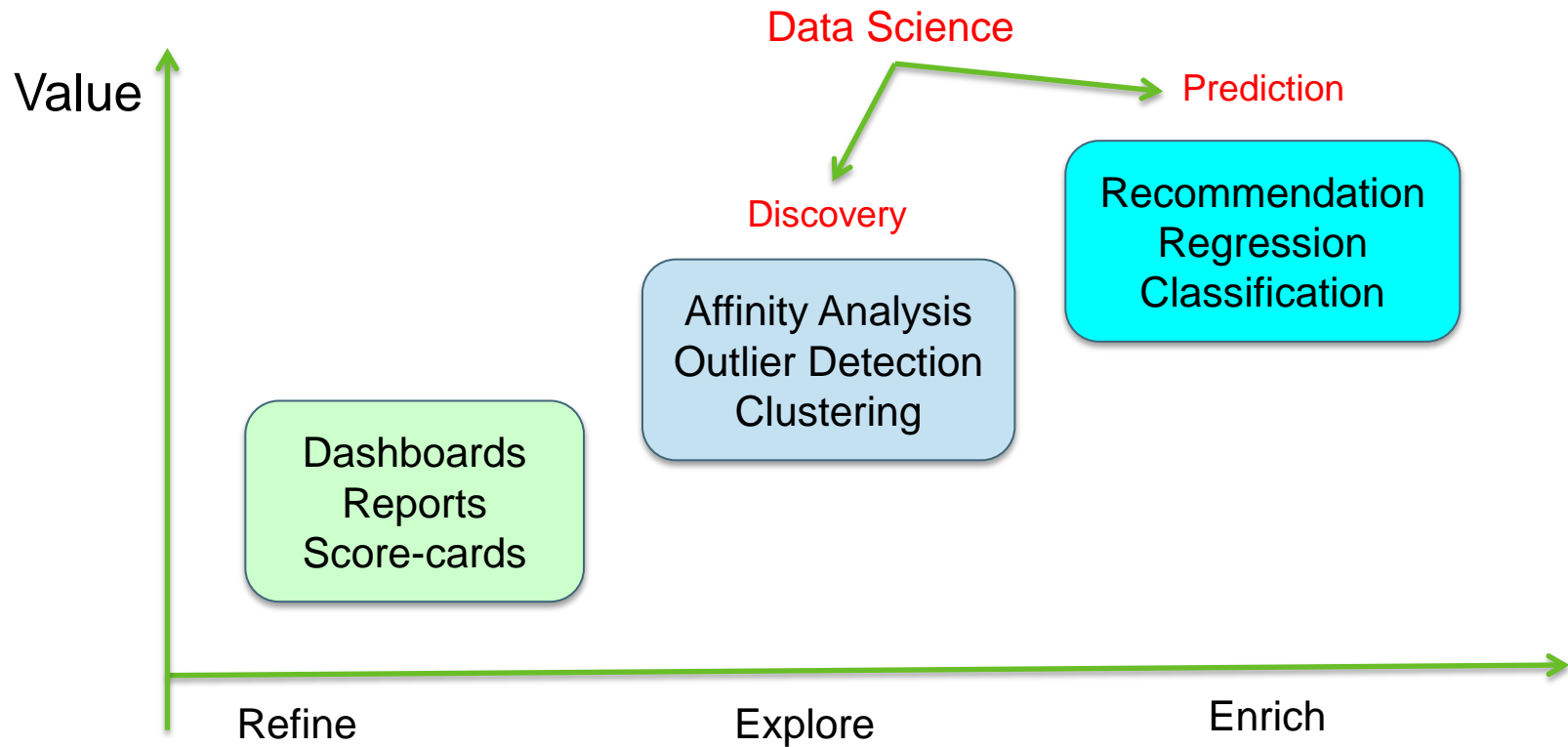
Web browser support



Mobile support



# Data science is a natural next step after business intelligence



**Business Intelligence:** measure & count; simple analytics

**Data Science:** discovery & prediction; complex analytics; “data product”

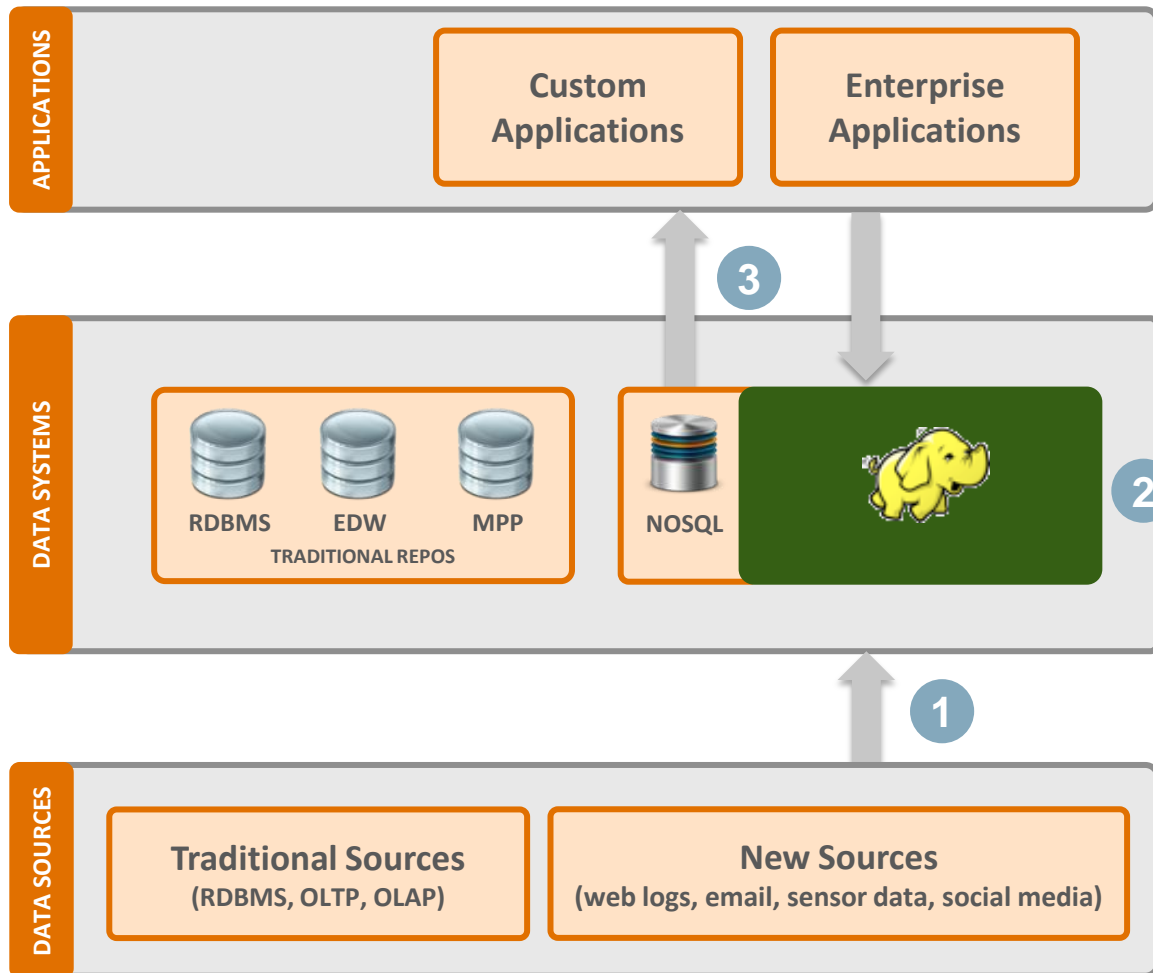
# Application Enrichment

# Application Enrichment

Refine

Explore

Enrich



Collect data, analyze and present salient results for online apps

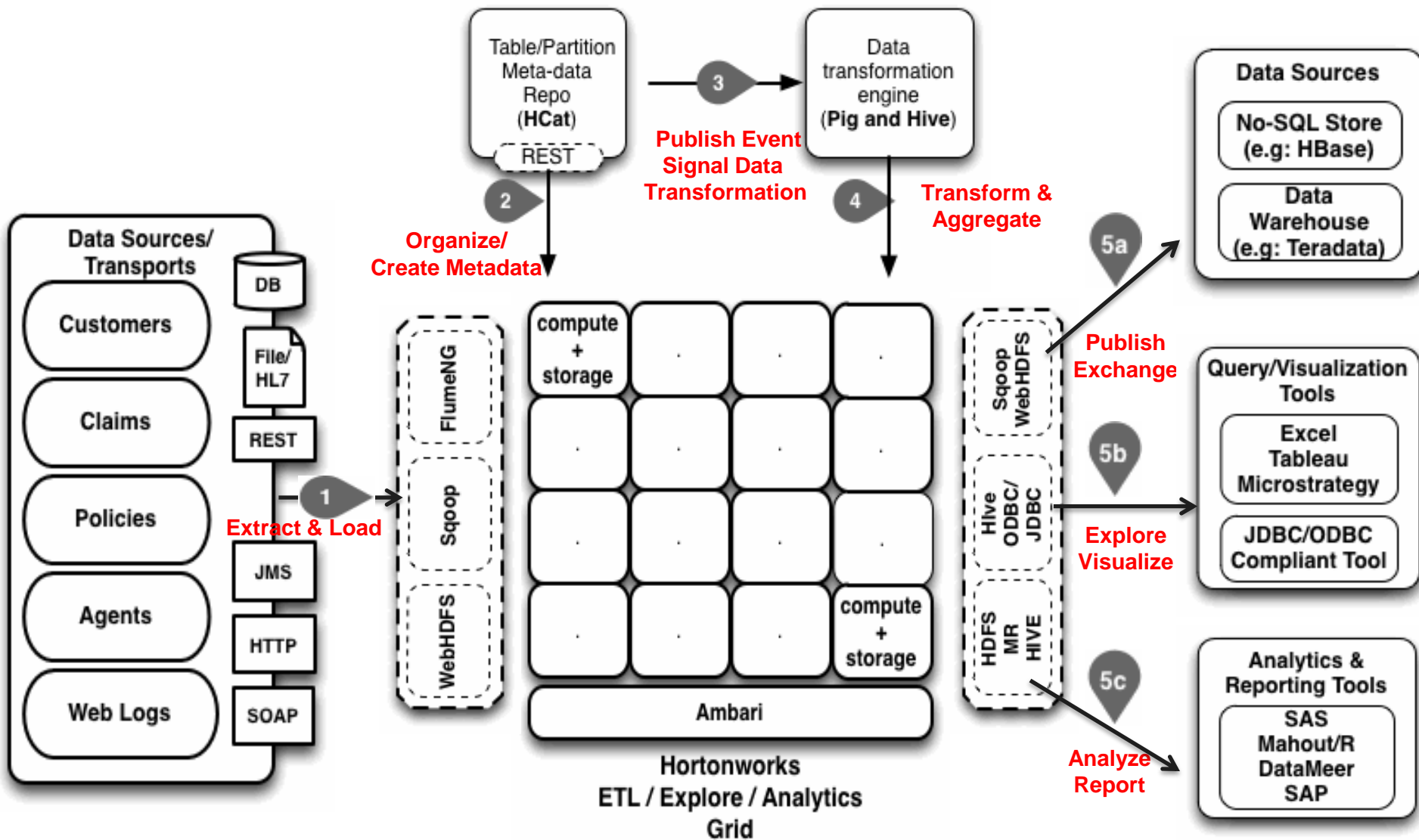
- 1 Capture**  
Capture all data
- 2 Process**  
Parse, cleanse, apply structure & transform
- 3 Exchange**  
Incorporate data directly into applications

# Key use-cases in Finance/Insurance

---

- Trading Analysis:
  - How do I predict Trading trends based on market sentiment?
  - How do I test Algorithms against years vs days fo data?
- Fraud detection:
  - Detect illegal credit card activity and alert bank/consumer
  - Detect illegal insurance claims
- Customer risk profiling:
  - How likely is this customer to pay back his mortgage?
  - How likely is this customer to get sick?
- Internal fraud detection (compliance):
  - Is this employee accessing financial information they are not allowed to access?

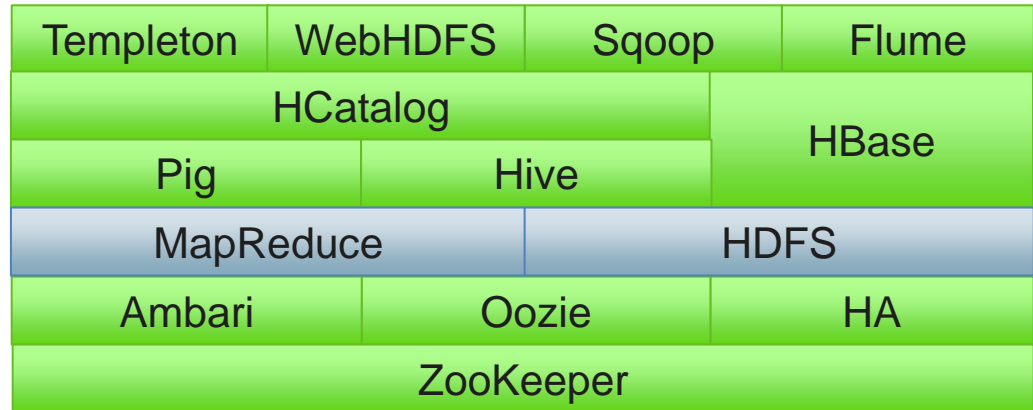
# HDP Reference Architecture



# From Community to the Enterprise

# What is a Hadoop “Distribution”

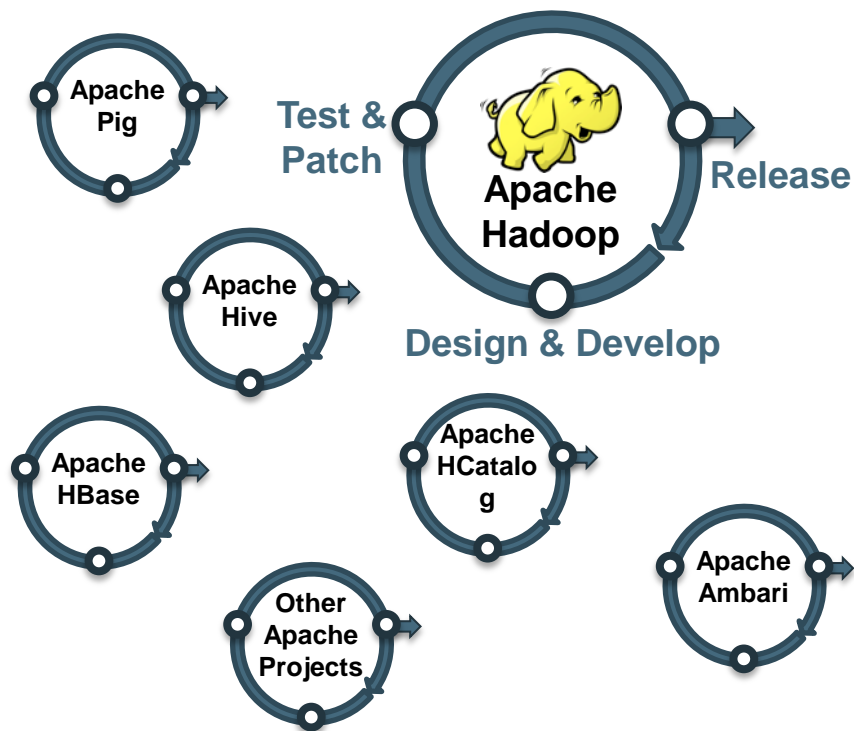
A complimentary set of open source technologies that make up a complete data platform



- 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 Community Leadership



*"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.."*

*- Jeff Kelly: Wikibon*

## Apache Software Foundation Guiding Principles

- Release early & often
- Transparency, respect, meritocracy

## Key Roles held by Hortonworkers

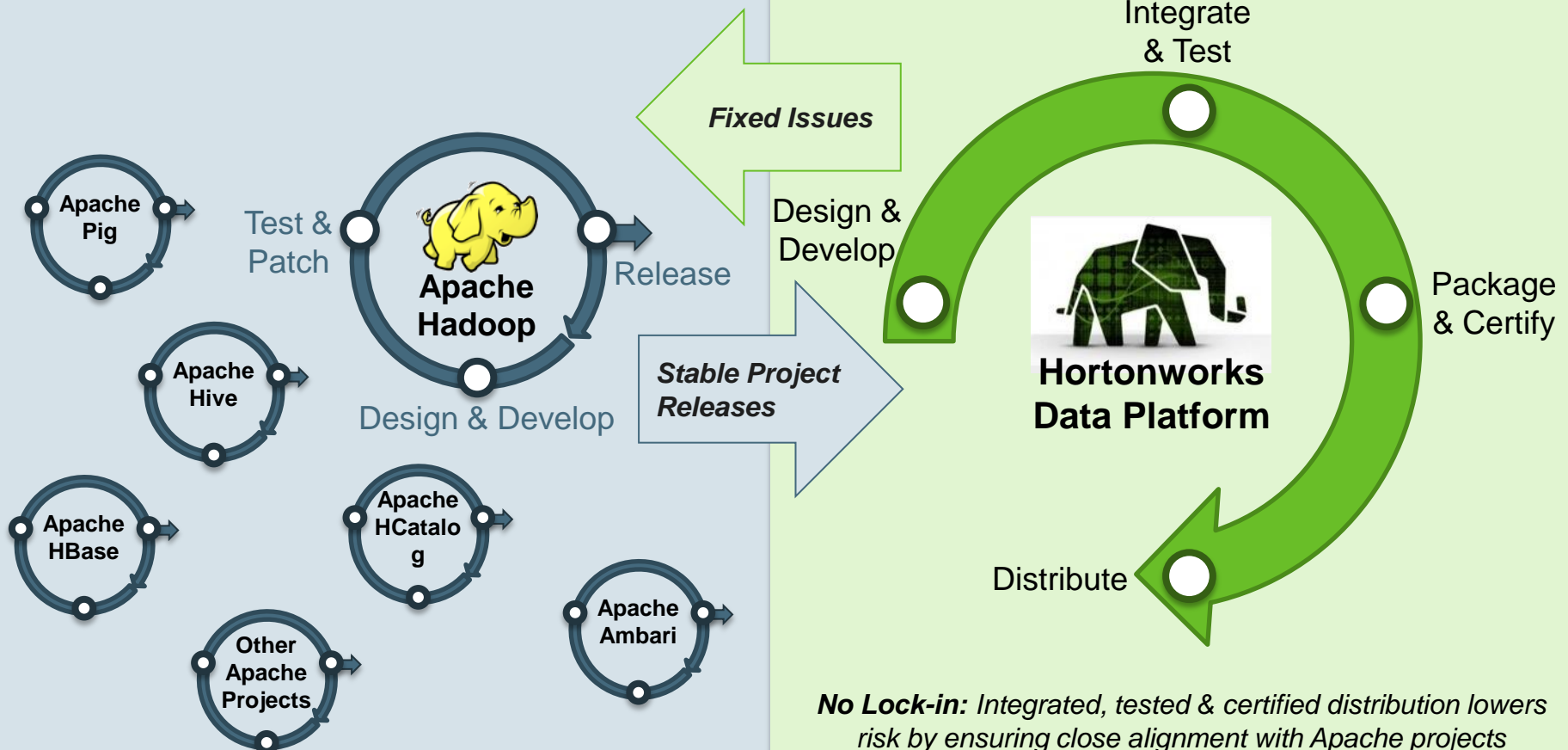
- VP & PMC Members
  - Arun Murthy (Hadoop), Daniel Dai (Pig), Mahadev Konar (Zookeeper)
- Release Managers
  - Matt Foley (Hadoop 1.x), Arun Murthy (Hadoop 2.x), Ashutosh Chauhan (Hive), Daniel Dai (Pig), Alan Gates (HCatalog), Mahadev Konar (Ambari)
- Committers (We can all be Contributors)
  - 54 across all Hadoop-related projects

# Hortonworks Process for Enterprise Hadoop

## Upstream Community Projects

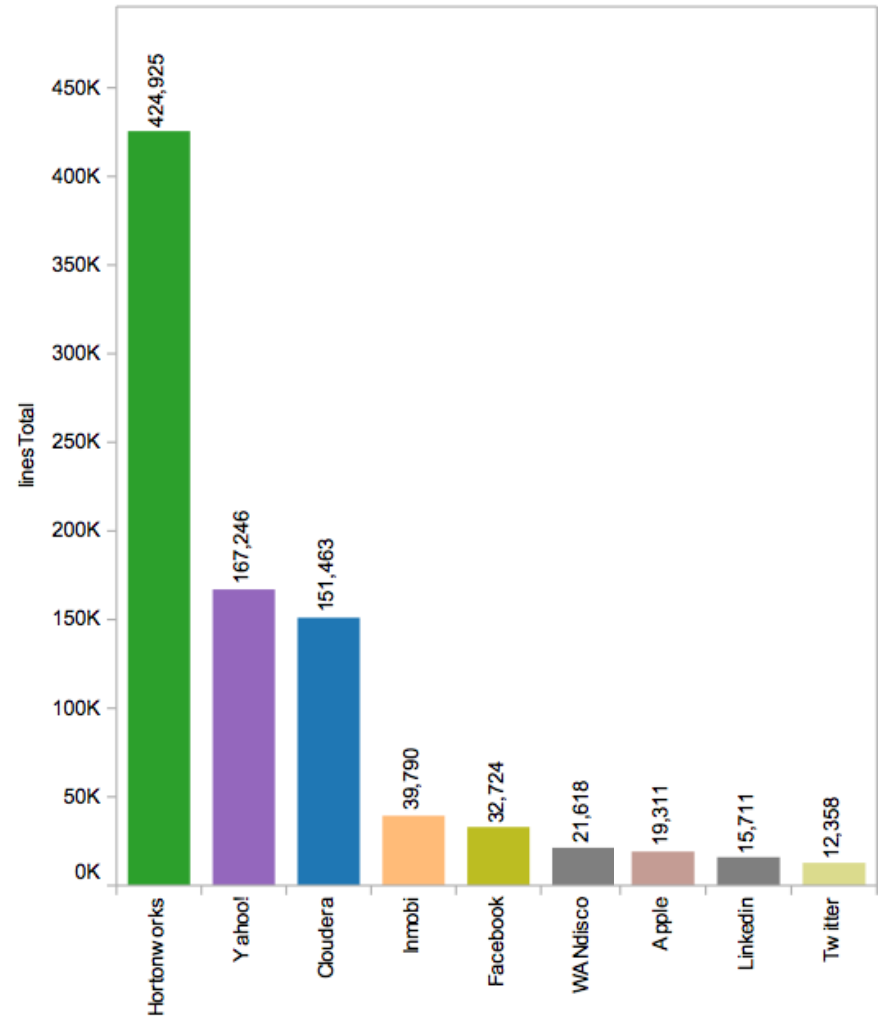
## Downstream Enterprise Product

*Virtuous cycle when development & fixed issues done upstream & stable project releases flow downstream*

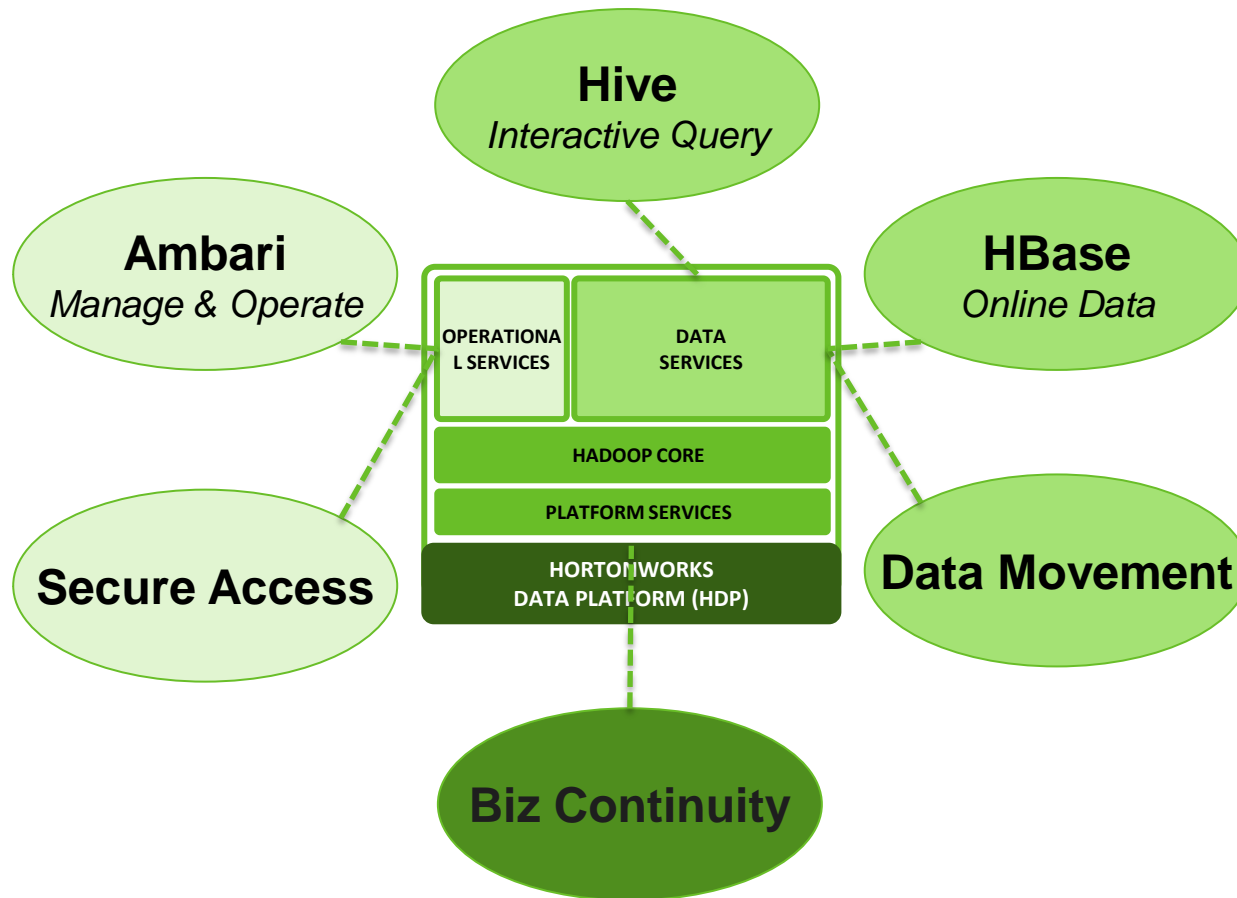


# Hadoop Evolution Starts at the Core

- **Over 1800 Contributors**
  - Very fast moving and improving
- **Driving next generation Hadoop**
  - YARN, MapReduce2, HDFS2, High Availability, Disaster Recovery
- **420k+ lines authored since 2006**
  - More than twice nearest contributor
- **Deeply integrating w/ecosystem**
  - Enabling new deployment platforms
    - (ex. Windows & Azure, Linux & VMware HA)
  - Creating deeply engineered solutions
    - (ex. Teradata big data appliance)
- **All Apache, NO holdbacks**
  - 100% of code contributed to Apache



# Where is Enterprise Apache Hadoop Going?



## –Platform Services

- Replication, Mirroring, Snapshots, ...

## –Data Services

- In support of Refine, Explore, Enrich

## –Operational Services

- Manageability, Security, ...

# Becoming Data Driven

# 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*



# Your Fastest On-ramp to Enterprise Hadoop™!



The Sandbox lets you experience Apache Hadoop from the convenience of your own laptop – no data center, no cloud and no internet connection needed!

The Hortonworks Sandbox is:

- A free download: <http://hortonworks.com/products/hortonworks-sandbox/>
- A complete, self contained virtual machine with Apache Hadoop pre-configured
- A personal, portable and standalone Hadoop environment
- A set of hands-on, step-by-step tutorials that allow you to learn and explore Hadoop

# Thank You!

## Questions & Answers

