Hortonworks

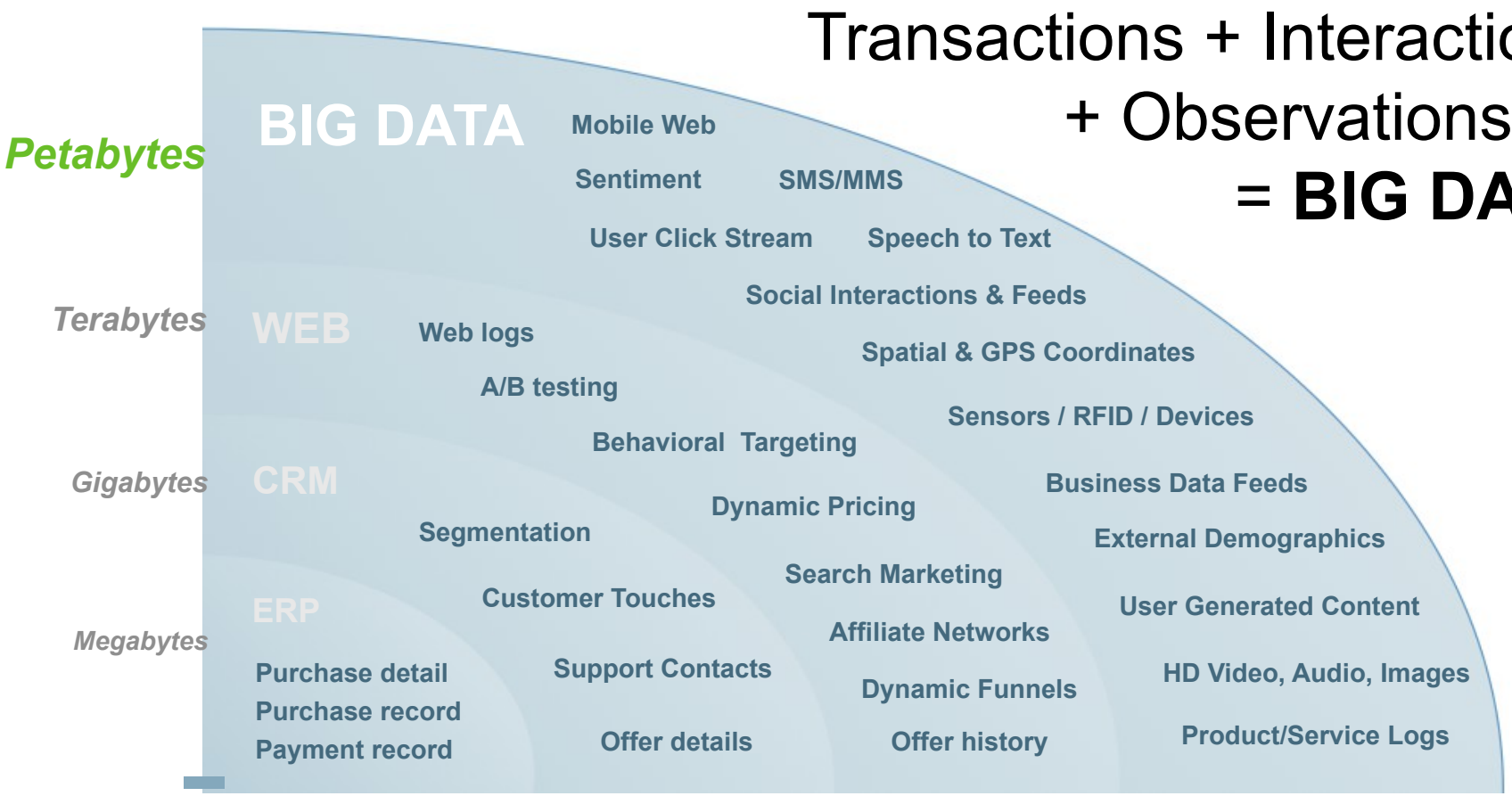
# Hadoop & Hortonworks

## Open Source Wild Fire

November 2012
OW2 Con

OW2con'12

# Big data changes the game



# Big Data: Optimize Outcomes at Scale

| | |
|---|---|
| **Sports** | Championships |
| **Intelligence** | Detection |
| **Finance** | Algorithms |
| **Advertising** | Performance |
| **Fraud** | Prevention |
| **Retail / Wholesale** | Inventory turns |
| **Manufacturing** | Supply chains |
| **Healthcare** | Patient outcomes |
| **Education** | Learning outcomes |
| **Government** | Citizen services |

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

# Apache Hadoop


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

Storage

**HDFS**


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

Processing

**Map Reduce**

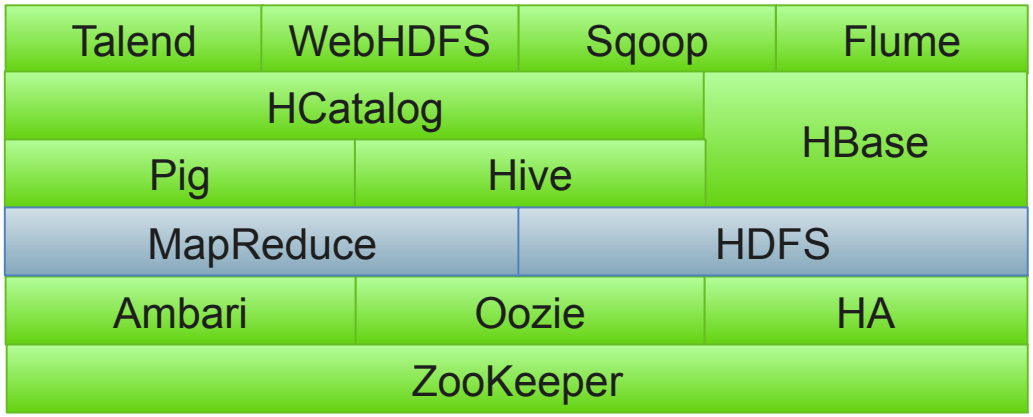

- Splits a task across processors “near” the data & assembles results
- Self-Healing, High Bandwidth Clustered Storage

## 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

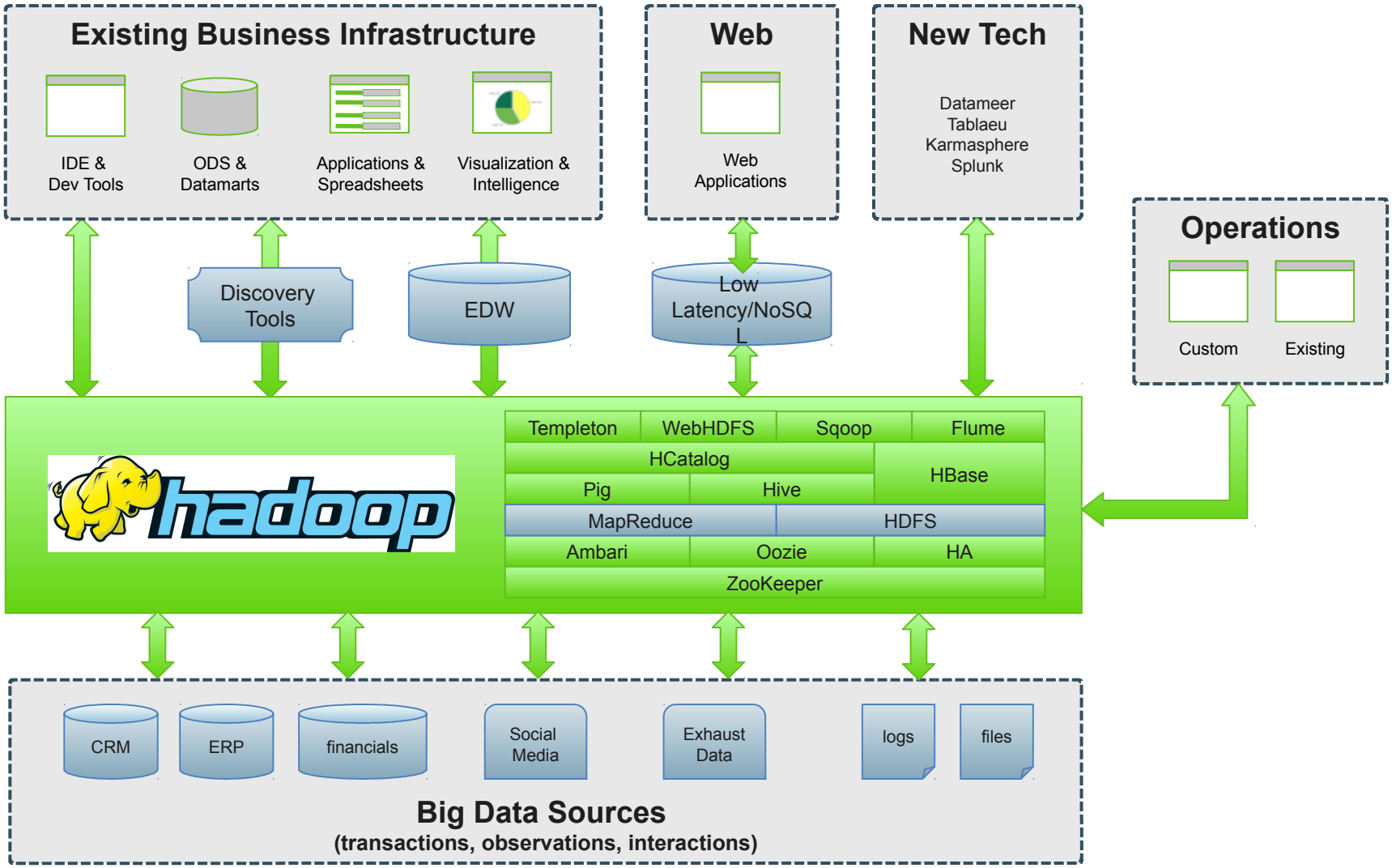
# What is a Hadoop “Distribution”

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

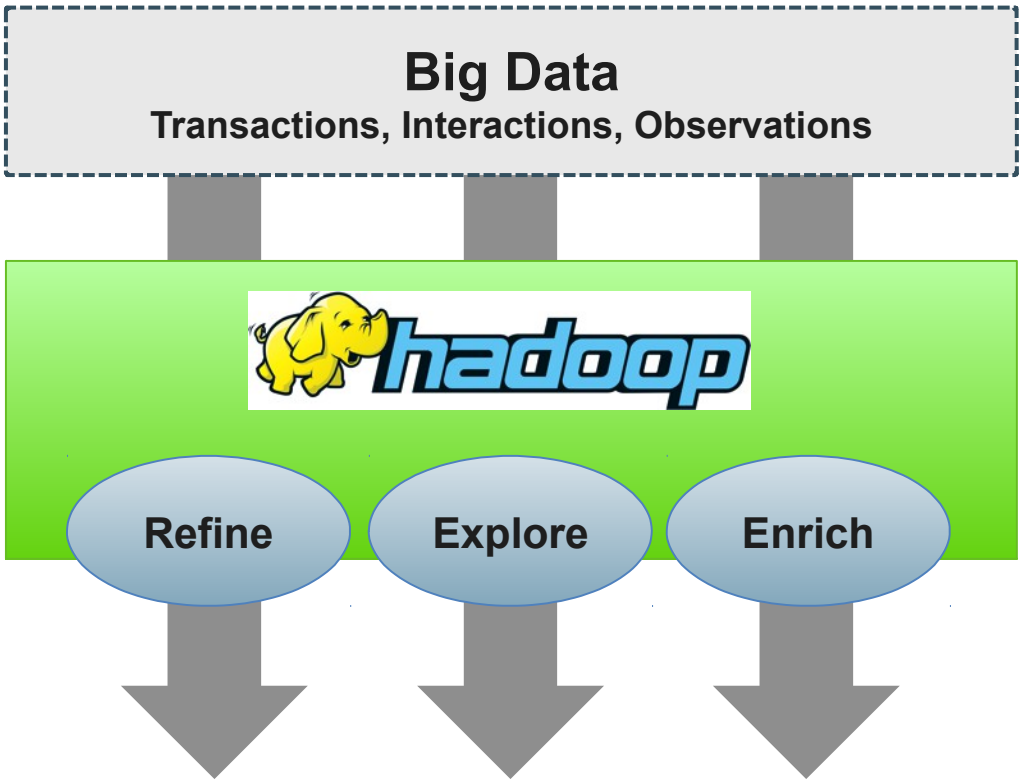



- Tested and pre-packaged to ease installation and usage
- Collects the right versions of the components that all have different release cycles and ensures they work together

# Hadoop in Enterprise Data Architectures



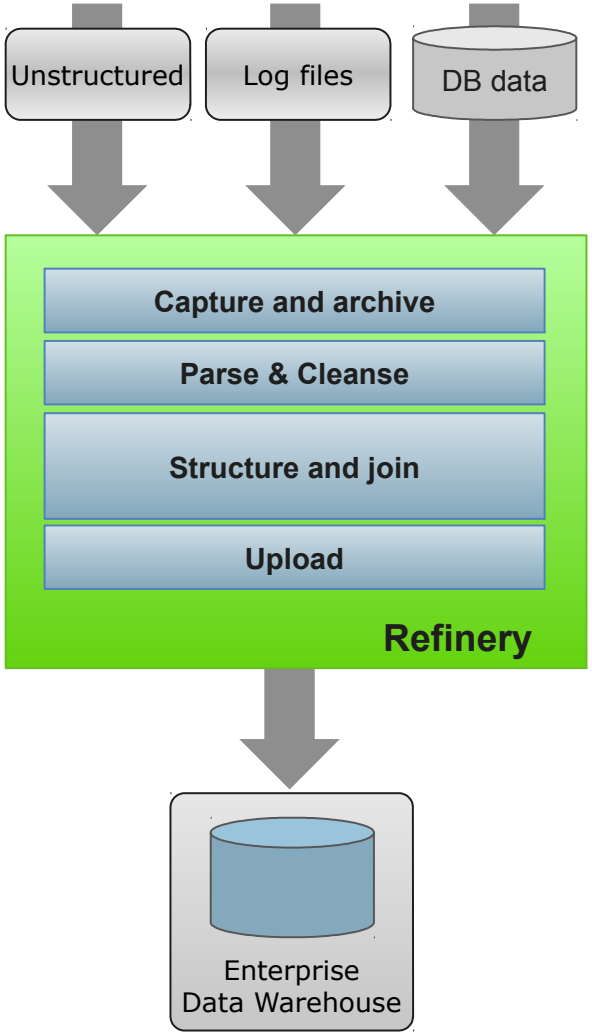
# Apache Hadoop & Big Data Use Cases



# 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

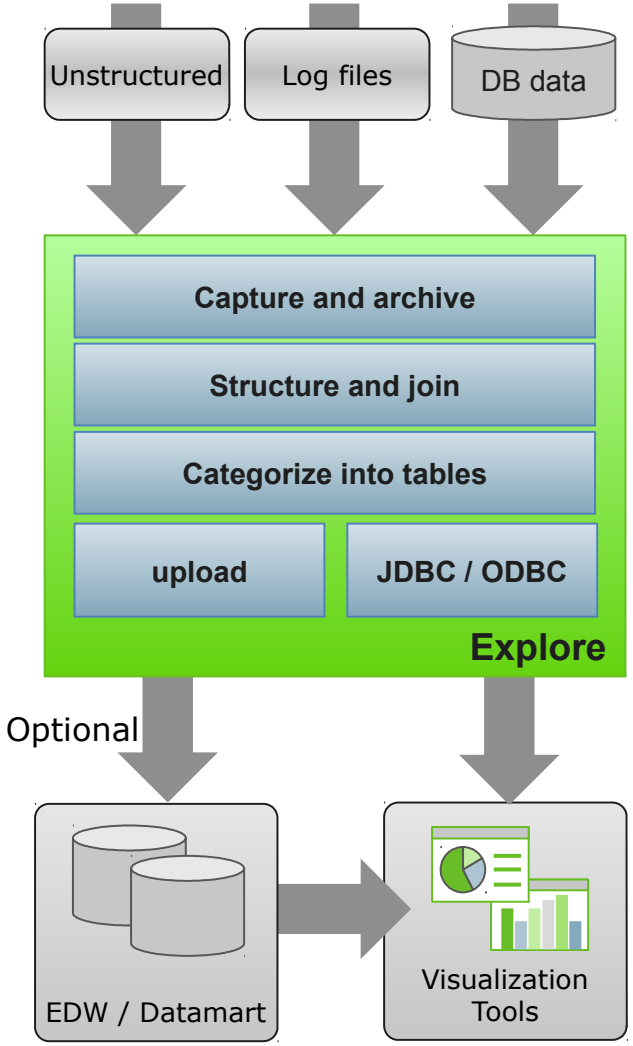
## Exchange

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

# Big 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

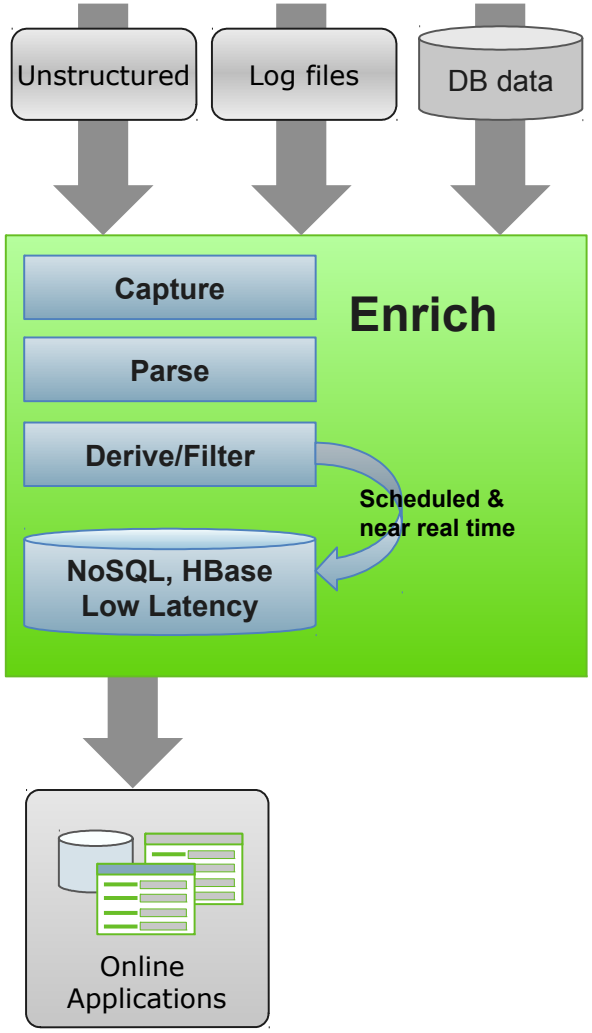
## Exchange

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

# Application Enrichment

*Deliver Hadoop analysis to online apps*

Refine | Explore | Enrich




## Capture

- Capture data that was once too bulky and unmanageable

## Process

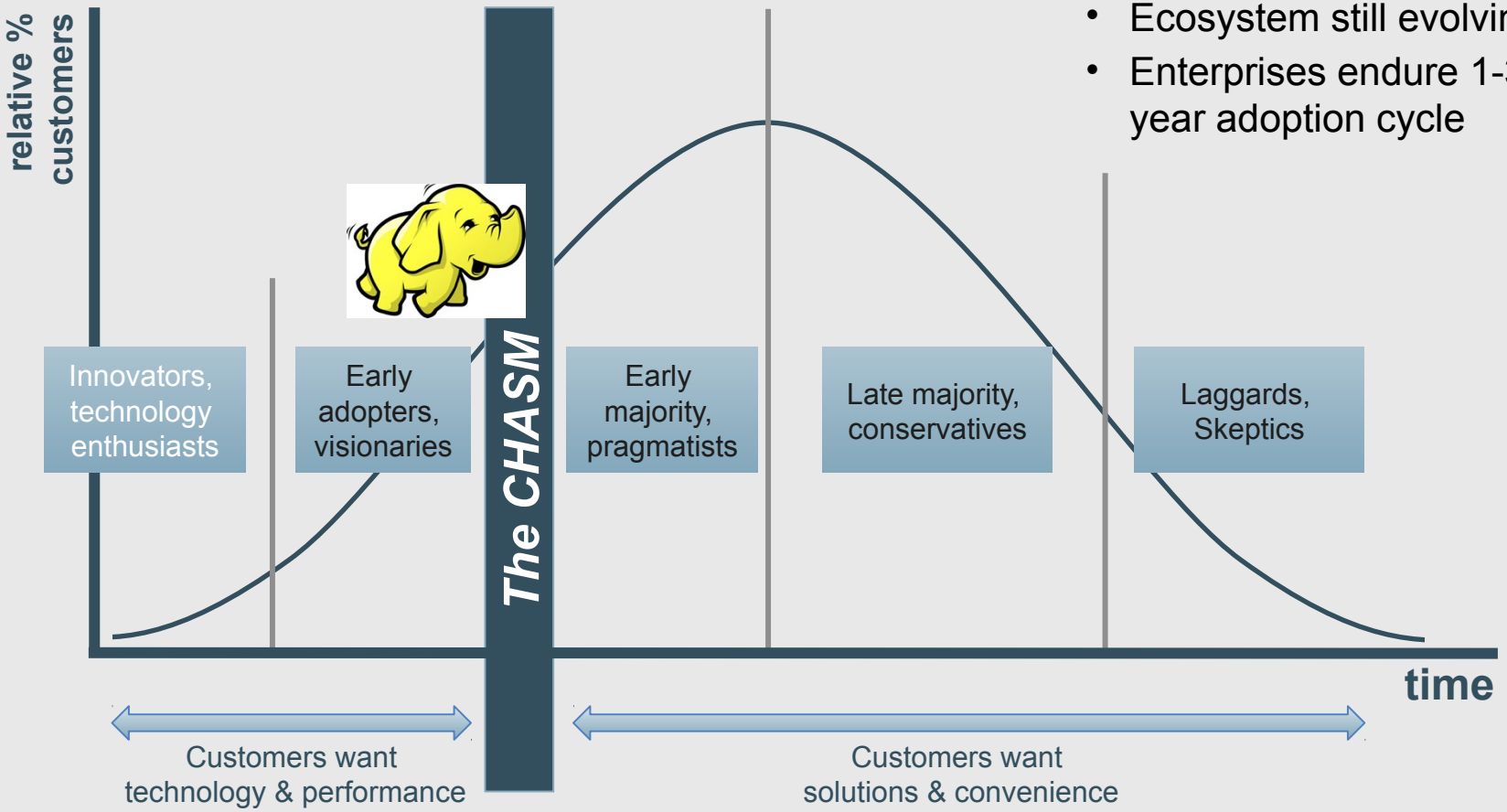
- Uncover aggregate characteristics across data
- Use Hive Pig and Map Reduce to identify patterns
- Filter useful data from mass streams (Pig)
- Micro or macro batch oriented schedules

## 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

# Balancing Innovation & Stability

- Hadoop is “pre-chasm”
- Ecosystem still evolving
- Enterprises endure 1-3 year adoption cycle




*Source: Geoffrey Moore - Crossing the Chasm*

# What Hortonworks does…


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

Strategy: invest in Apache Hadoop to make it ***"The enterprise big data platform"***

## Distribution

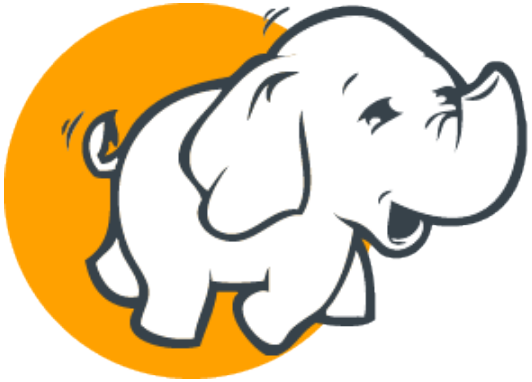
- **Hortonworks Data Platform (HDP)**
- Enterprise Ready, Stable, Reliable, Tested
- 100% open source
- Built by the architects, builders and operators of Apache Hadoop

## Ecosystem

- **Enable an Ecosystem of Big Data Apps**
- Our goal os to make sure all your tools work WITH Hadoop
- HDP is Hadoop for
  - Microsoft
  - Teradata

## Support

- **Deliver highest quality support and expertise**
- Access to Apache Hadoop Experts
- Hadoop training an certification by the Hadoop experts(web, public, private)



2013 HADOOP SUMMIT

## AMSTERDAM

## March 20-21, 2013

### Enabling the Next Generation Enterprise Data Platform

- LEARN: Dozens of Sessions
- INTERACT: Community Focused Event

**Register today! @ hadoopsummit.org**

Hortonworks

OW2con'12