Hadoop & Hortonworks

Open Source Wild Fire

November 2012
OW2 Con
Big data changes the game

Transactions + Interactions + Observations = BIG DATA

Increasing Data Variety and Complexity
Big Data: Optimize Outcomes at Scale

- Sports
- Intelligence
- Finance
- Advertising
- Fraud
- Retail / Wholesale
- Manufacturing
- Healthcare
- Education
- Government

- Championships
- Detection
- Algorithms
- Performance
- Prevention
- Inventory turns
- Supply chains
- Patient outcomes
- Learning outcomes
- Citizen services

Apache Hadoop

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

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

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

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

Existing Business Infrastructure
- IDE & Dev Tools
- ODS & Datamarts
- Applications & Spreadsheets
- Visualization & Intelligence

Web
- Discovery Tools
- EDW
- Low Latency/NoSQ
- Web Applications

New Tech
- DataMeer
- Tableau
- Karmasphere
- Splunk

Operations
- Custom
- Existing

Big Data Sources
(transactions, observations, interactions)

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

CRM
ERP
financials
Social Media
Exhaust Data
logs
files

© Hortonworks Inc. 2012
Apache Hadoop & Big Data Use Cases

Big Data
Transactions, Interactions, Observations

Refine
Explore
Enrich

Business Case
Operational Data Refinery
*Hadoop as platform for ETL modernization*

**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_

**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

**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/offers 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

Customers want technology & performance
Innovators, technology enthusiasts
Early adopters, visionaries

Customers want solutions & convenience
Early majority, pragmatists
Late majority, conservatives
Laggards, Skeptics

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)
AMSTERDAM
March 20-21, 2013

Enabling the Next Generation
Enterprise Data Platform

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

Register today!  @ hadoopsummit.org