Intel IT Center

Apache Hadoop Framework
The Nexus of Open Source Innovation

Eric Baldeschwieler, CTO, Hortonworks
Avik Dey, Director, Hadoop Services, Intel

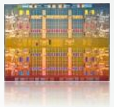
Moderator: Todd Cramer, Director Product Marketing, Intel

Hortonworks

intel

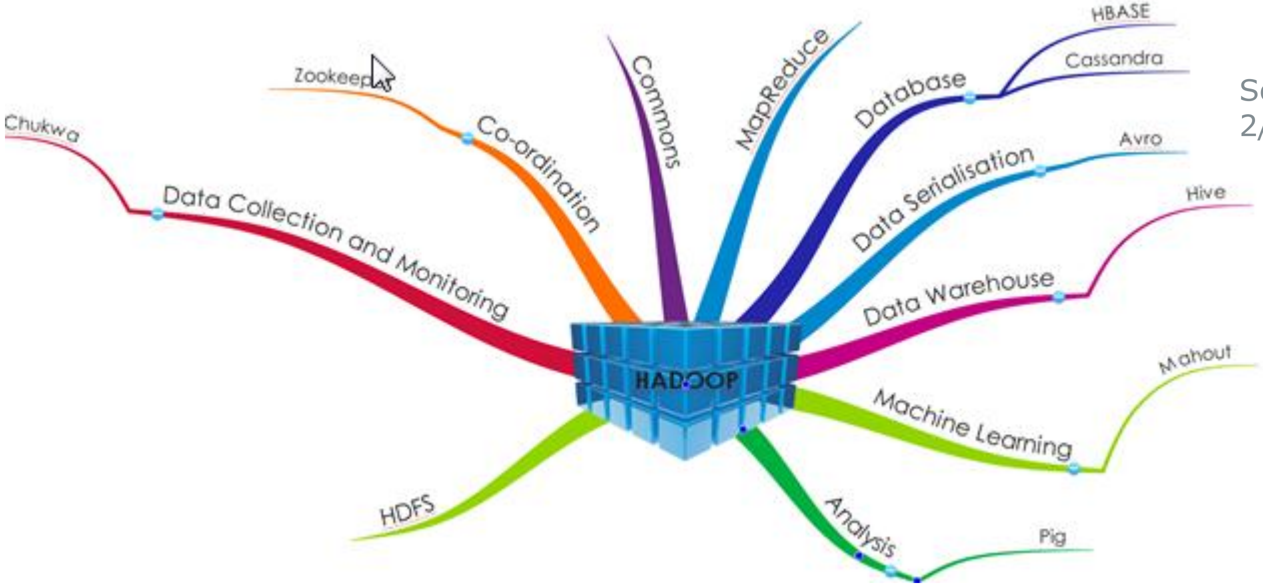
Evolution to Open Source Data Management with Scale-out Storage & Processing

Intel IT Center

Date	Paradigm	Processing Style/ Scale Out	Form Factor
90s	• Reporting / Data Mining • High Cost / Isolated use	• Batch– "sales reports" • Sequential SQL queries Scale  Multi-core	RDBMS
2000s	• Model-based discovery • High Cost / Dept Use	• Batch-ie correlated buying pattern • No SQL. parallel analysis • Shared disk/memory Scale Node Node Node	No SQL RDBMS Proprietary MPP/ DW Appliance
Today	• Unbounded Map Reduce Query • Low Cost / Enterprise Use • Arrival of vast amounts of unstructured data	• Real-time- ie recommend engine • Process @ storage node • Built-in data replication/reliability • Shared nothing, in memory Unlimited Linear Scale Distributed node addition	Open Source SW coupled to commodity HW Node Node Node hadoop

Apache Hadoop Evolution

Intel IT Center



Source - Steven Nimmons 2/24-12

2006

- HDFS
- MapReduce

2008

- HBase
- ZooKeeper
- Pig
- Hive

2009-10

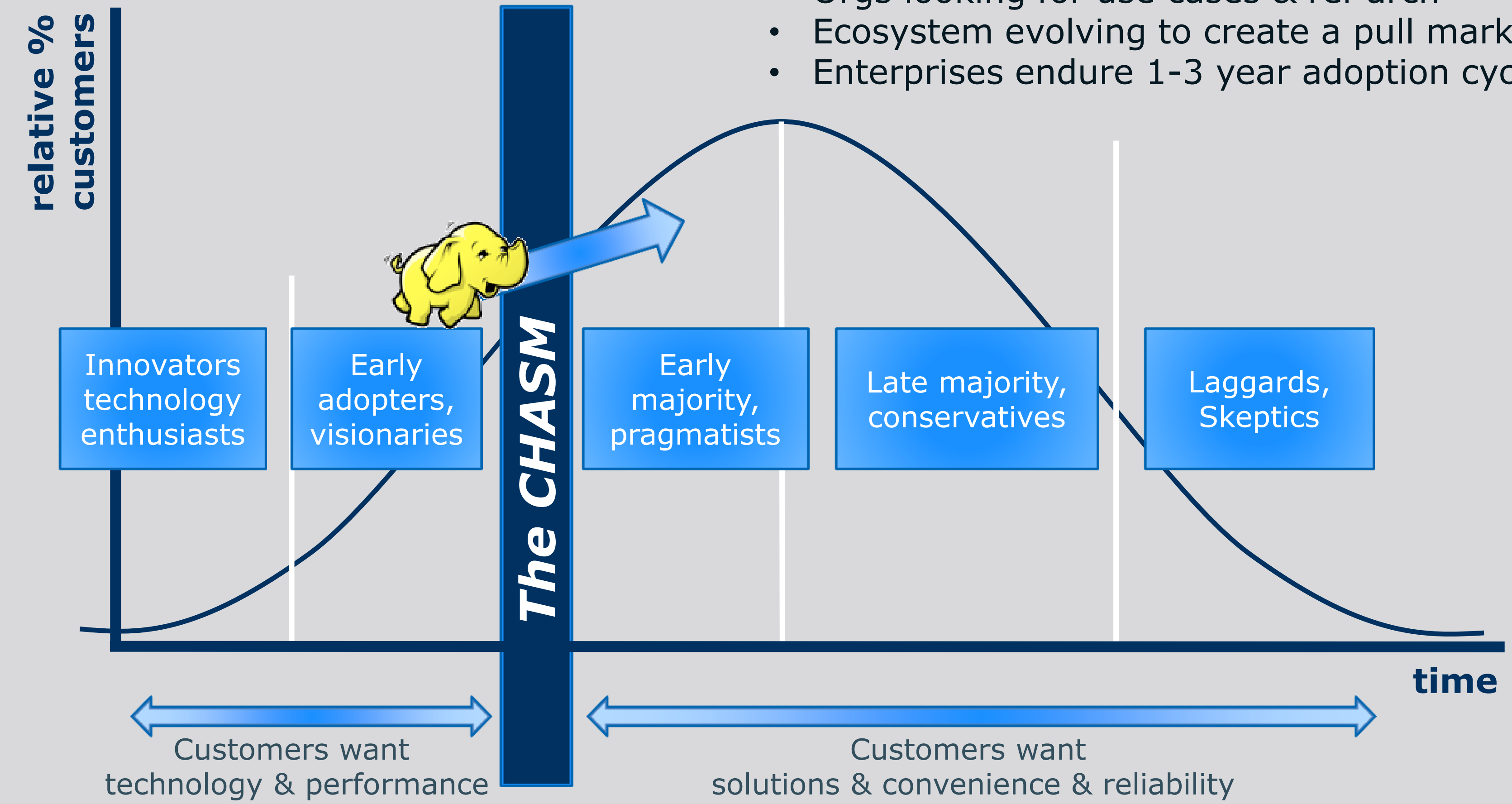
- Flume
- Avro
- Whirr
- Sqoop
- Mahout
- Oozie

2011-12

- HCatalog
- Bigtop
- Ambari
- Yarn

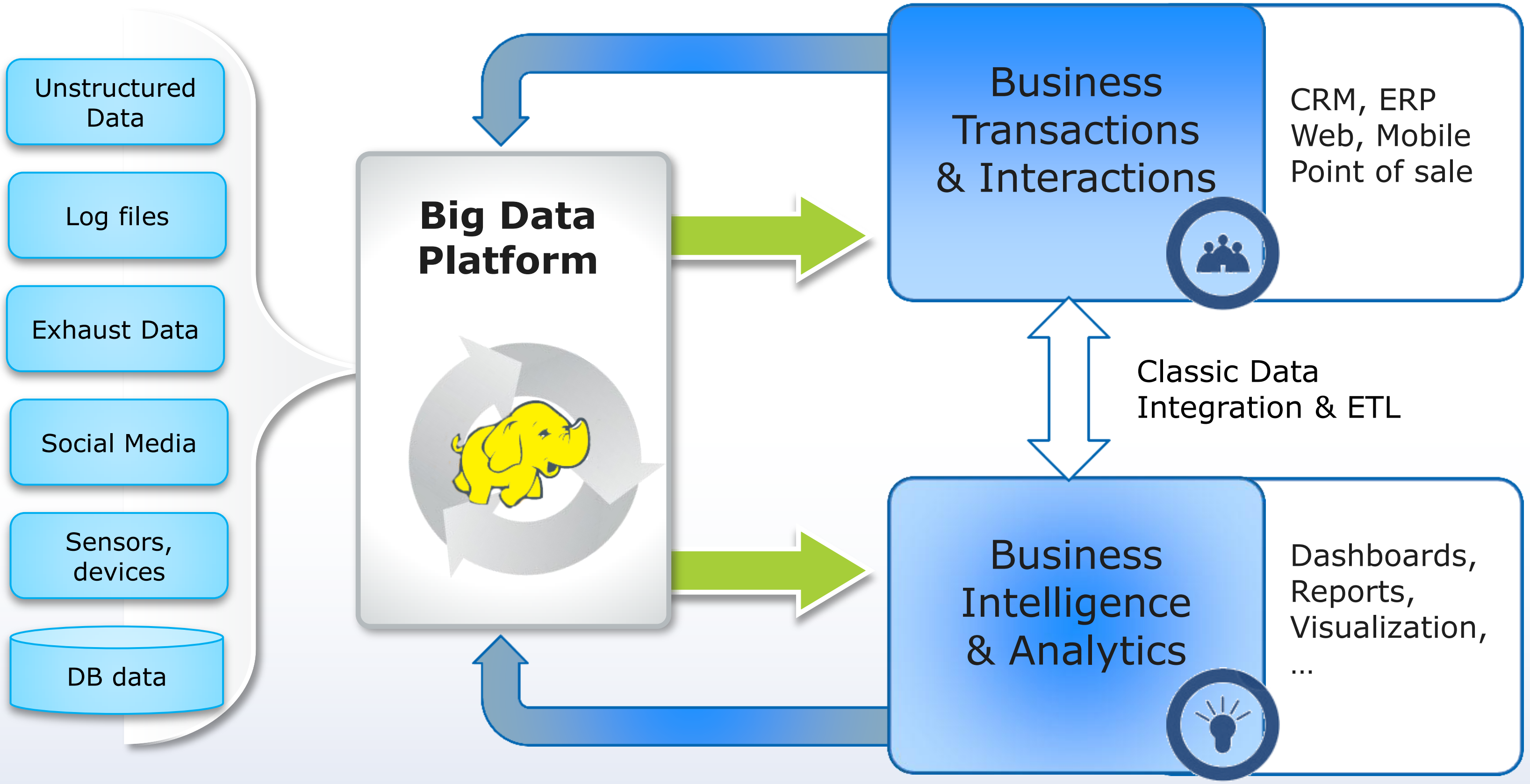
Hadoop: What will it take to cross The Chasm?

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


relative % customers
Innovators technology enthusiasts
Early adopters, visionaries
The CHASM
Early majority, pragmatists
Late majority, conservatives
Laggards, Skeptics
time
Customers want technology & performance
Customers want solutions & convenience & reliability

Source: Geoffrey Moore - Crossing the Chasm

Enterprise Big Data Flows



Unstructured Data

Log files

Exhaust Data

Social Media

Sensors, devices

DB data

Big Data Platform

Business Transactions & Interactions

CRM, ERP
Web, Mobile
Point of sale

Classic Data Integration & ETL

Business Intelligence & Analytics

Dashboards,
Reports,
Visualization,
…

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

2 **Process**
Transform, refine, aggregate, analyze, report

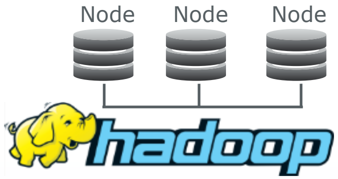
3 **Exchange**
Interoperate and share data with applications/analytics

What changes from POC to large clusters?

5-100 nodes
“Small cluster”

4000 node
“Hadoop at Scale”

Cluster Size

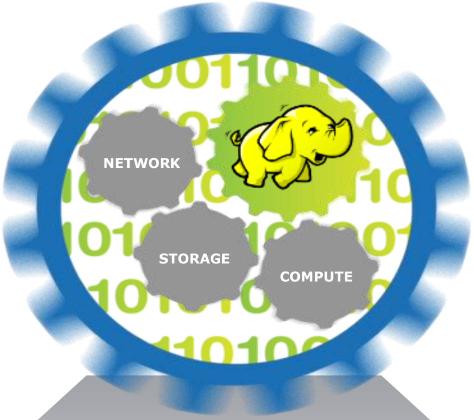


- Staff & consultants are dominant costs
- Redundant networks, hardware reliability features save human capital & support
- Need to focus on simplicity

- Hardware + Power + Hosting are dominant costs
- Hardware Optimization
- Failures are inevitable, Hadoop software handles this
- Hadoop operations expertise

Optimizing Hadoop Deployments

Address Potential Deployment Bottlenecks



Benchmark Tuning	Security & APIs	Compute	Disk Write/memory	Fast Fabric
Hi-tune Hi-Bench	Encryption	Instruction Sets	SSDs Non-volatile memory	10GbE

Talk to an Expert: Question & Answer

Today's Experts:

- Eric Baldeschwieler, CTO, Hortonworks - @JERIC14
- Avik Dey, Director, Hadoop Services, Intel - @AvikonHadoop

Submit your questions:

- Ask questions at anytime by pressing the Question tab at the top of the player.

Download today's content:

- Located under the attachment tab at the top of the player

More information:

- www.intel.com/bigdata

intel®