(intel') Look Inside:
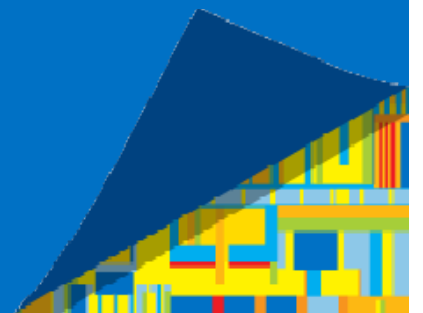
Intel® HPC Distribution for Apache Hadoop* Software including Intel® Enterprise Edition for Lustre* Software

SC13, November, 2013

Agenda

- **Abstract**

- **Opportunity:**

  - HPC Adoption of Big Data Analytics on Apache Hadoop*

  - Enterprise Adoption of Technical Computing on HPC Systems

- **Challenge:** Need for an Efficient Infrastructure for Hadoop and HPC Workloads

- **Solution:** Intel® HPC Distribution for Apache Hadoop* Software

  - **Architecture:** Key Differentiators

- **Value Prop:** Features, Functions, and Benefits

- **Proof Points**

- **Intel HPC Distribution** BETA PROGRAM

Abstract

Intel is addressing the need for a scalable, efficient infrastructure that can support Big Data analytics applications on HPC systems in the enterprise by introducing the Intel® HPC Distribution product bundle and inviting early adopters to a BETA program.

Opportunity: HPC Adoption of Big Data Analytics on Apache Hadoop*

Discoveries and Decisions Driven by Big Data Analytics

Operational Efficiency

- Traffic Optimization
- Smart Energy Grid

Consumer Behavior

- Location Aware Ad Placement
- Buyer Protection Program

Security and Risk Management

- Personalized Preventive Care
- Claim Fraud Reduction

Opportunity: Enterprise Adoption of Technical Computing on HPC Systems

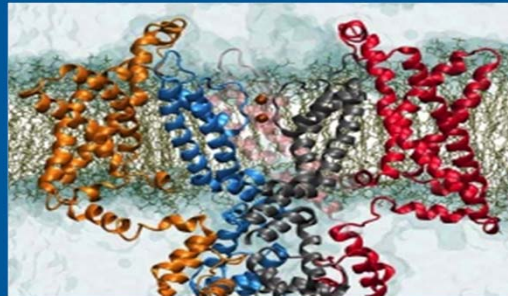## Discoveries and Decisions Driven by Big, Fast Supercomputers

Geosciences	Life Sciences	Large scale manufacturing
▪ Oil and gas exploration ▪ Seismic modeling ▪ Modeling wind turbine placement	▪ Genomics ▪ Drug discovery	▪ Crash safety for auto and aerospace ▪ Virtual prototype

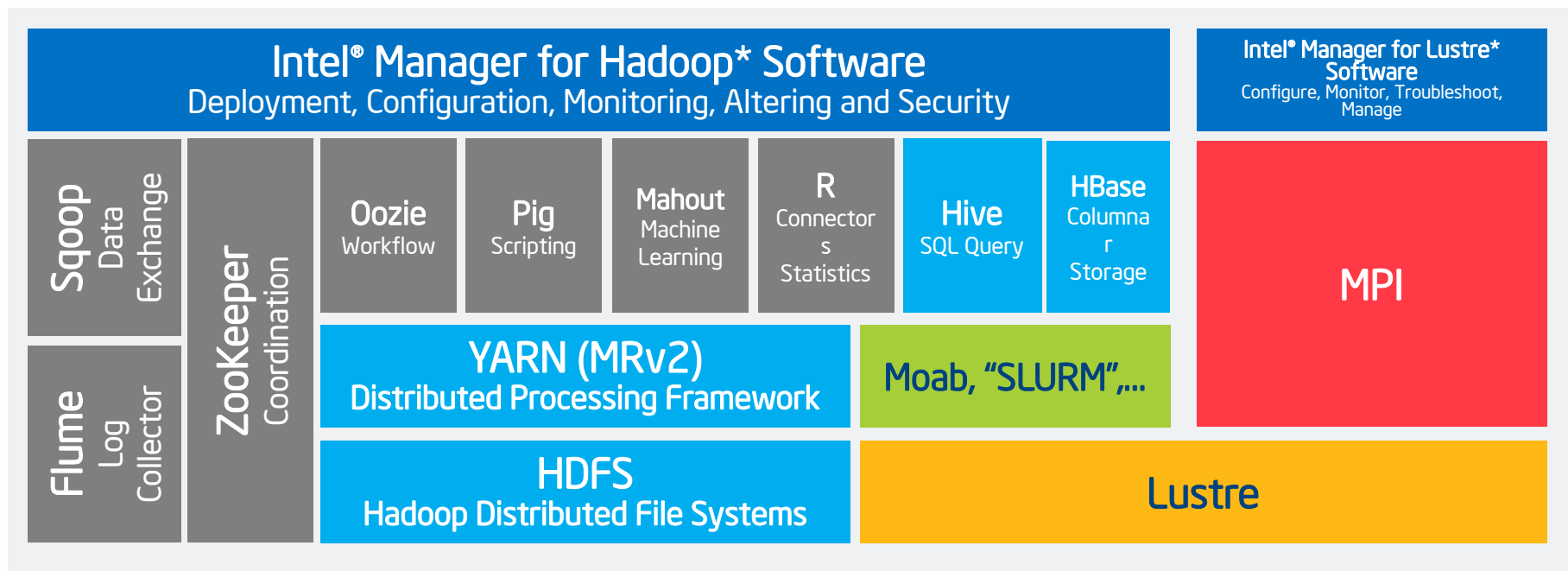# Challenge: Need for an Efficient Infrastructure for Hadoop* and HPC Workloads

Tackling a wide range of previously intractable problems that are important for economic competitiveness, scientific advancement, national security, and the quality of human life.

These include fraud detection, antiterrorist analysis, social and biological network analysis, semantic analysis, financial and economic modeling, drug discovery and epidemiology, weather and climate modeling, oil exploration, and power grid management.

The common denominator is that the problems are large and complex enough to require modeling and simulation on HPC resources.

Addressing the HPC Big Data Challenge
Intel® HPC Distribution for Apache Hadoop* Software

Intel® Manager for Hadoop* Software
Deployment, Configuration, Monitoring, Altering and Security

Intel® Manager for Lustre* Software
Configure, Monitor, Troubleshoot, Manage

Sqoop Data Exchange

Flume Log Collector

ZooKeeper Coordination

Oozie Workflow

Pig Scripting

Mahout Machine Learning

R Connectors Statistics

Hive SQL Query

HBase Columnar Storage

MPI

YARN (MRv2) Distributed Processing Framework

Moab, "SLURM",...

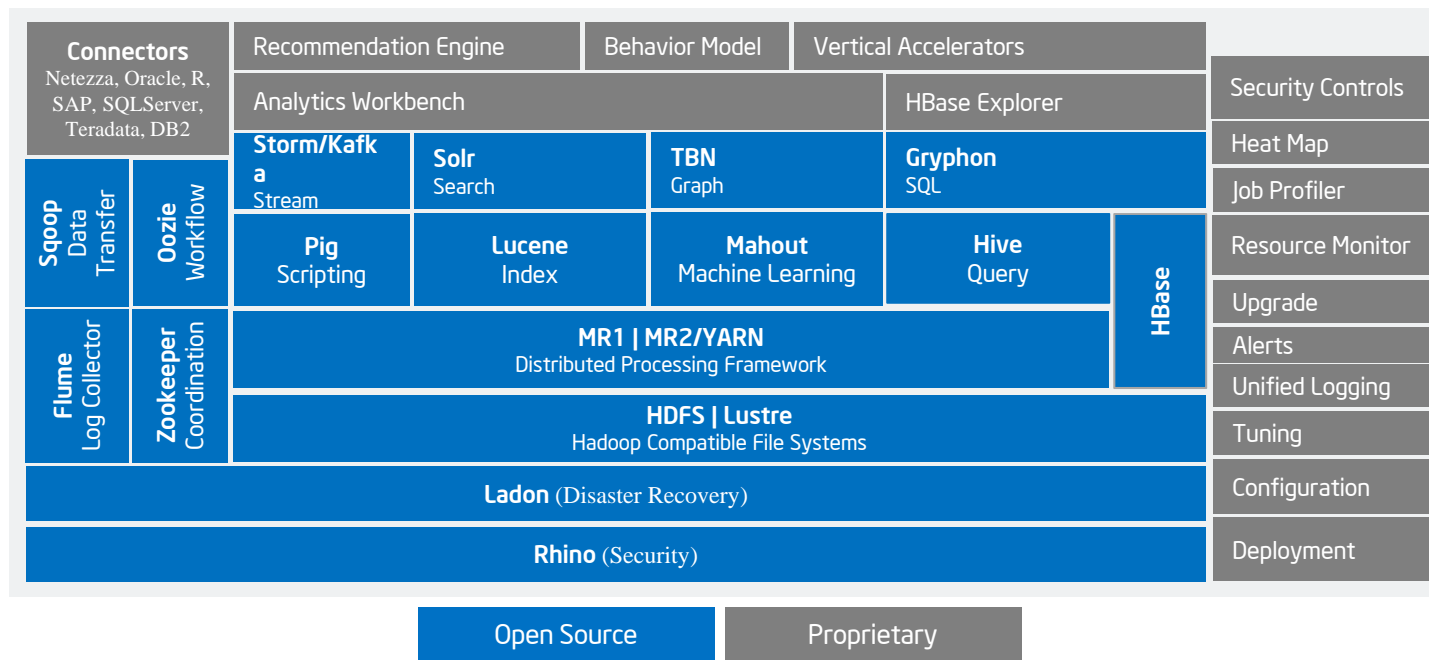**HDFS** Hadoop Distributed File Systems

Lustre

Solution: Intel® HPC Distribution for Apache Hadoop* Software

Intel is the first to offer Lustre's parallel file system integrated with Hadoop workloads.

Intel® Distribution for Apache Hadoop* Software

- Authentication, authorization, auditing built-in to Apache Hadoop
- Transparent encryption in Hive, Pig, MapReduce, HBase, HDFS
- Up to 20x faster en/decryption with Intel AES-NI[1]
- Up to 30x faster on Intel architecture than other hardware

- Up to 2.6X faster than other open source distributions
- Enterprise-grade Hadoop cluster management console and APIs
- Automated configuration with Intel® Active Tuner
- Direct integration to Intel EE for Lustre allows users to utilize that file system in place of the Hadoop Distributed File System (HDFS).

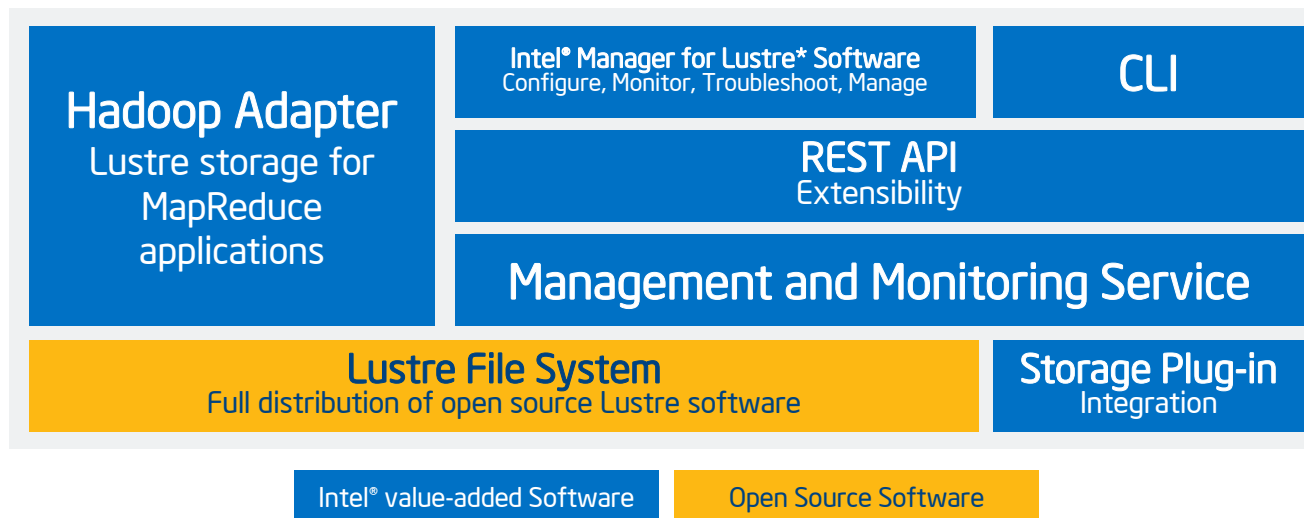| Connectors Netezza, Oracle, R, SAP, SQLServer, Teradata, DB2	Recommendation Engine		Behavior Model	Vertical Accelerators		Security Controls
	Analytics Workbench			HBase Explorer		Heat Map
Sqoop Data Transfer / Oozie Workflow	Storm/Kafka Stream	Solr Search	TBN Graph	Gryphon SQL		Job Profiler
	Pig Scripting	Lucene Index	Mahout Machine Learning	Hive Query	HBase	Resource Monitor
Flume Log Collector / Zookeeper Coordination	MR1 | MR2/YARN Distributed Processing Framework					Upgrade
						Alerts
	HDFS | Lustre Hadoop Compatible File Systems					Unified Logging
						Tuning
Ladon (Disaster Recovery)						Configuration
Rhino (Security)						Deployment

Open Source   Proprietary

Solution: Intel® Enterprise Edition for Lustre* Software

Intel® HPC Distribution for Apache Hadoop software is the only distribution of Apache Hadoop* to integrate and support Lustre* out of the box.

Intel® Enterprise Edition for Lustre* Software

- Full open source core
- Simple GUI for install and management with central data collection
- Direct integration with storage HW and applications
- Global tier-1 support

- Storage plug-in; deep vendor integration
- REST API—extensibility
- Hadoop* Adapter for shared simplified storage for Hadoop

Hadoop Adapter Lustre storage for MapReduce applications	Intel® Manager for Lustre* Software Configure, Monitor, Troubleshoot, Manage	CLI
	REST API Extensibility	
	Management and Monitoring Service	

Lustre File System Full distribution of open source Lustre software	Storage Plug-in Integration

Intel® value-added Software | Open Source Software

Intel® HPC Distribution: Open Platform for High Performance Data Analytics
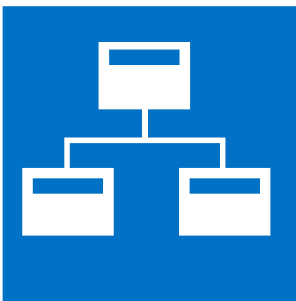
Value Prop: Features, Functions, and Benefits

Performance

- Bring compute to the data: Run MapReduce* on Lustre* without code changes
- Run MapReduce* faster: Avoid the intermediate file shuffle with shared storage

Efficiency

- Avoid Hadoop* islands in the sea of HPC systems
- Run MapReduce jobs alongside HPC workloads with full access to the cluster resources

Manageability

- Use the seamless integration to manage one common platform for Hadoop and HPC
- Develop with multiple programming models and deploy on shared storage

Proof Points

In IDC's 2013 worldwide study of HPC end users, 67% of the sites said they perform HPDA on their HPC systems, often using Hadoop*, with an average of 30% of the available computing cycles devoted to this work.

This formative market for Big Data problems needing HPC includes data-intensive modeling and simulation, along with newer analytics methods employed by established HPC users and first-time users from the commercial world.
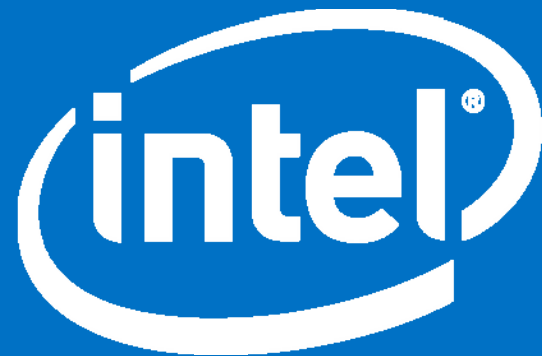
Join the BETA program

Early adopters of the combined Intel Distribution for Apache Hadoop Software and Intel EE for Lustre Software solution will receive a free, exclusive limited-use version of the software and exchange insights with Intel experts.

To be considered for the BETA, please contact:

- hpdd-info@intel.com