Vendor Spotlight
Cloudera Opens Up the World of Big Data
Omer Trajman, Vice President of Technology Solutions, Cloudera

Omer Trajman talks about how Cloudera is making major contributions to Apache open-source projects, and how Cloudera’s own Hadoop distribution is helping organizations profit from all their data.

Welcome to the New IT Model

If you look at the past 10 or 15 years of IT, it’s been a continual refinement of a pretty standard model: storage over here, computer over there, and relational databases, applications, and Web tiers on top. Big data doesn’t fit into that model. Big data necessitates new technologies, and that requires new processes. It’s a brand-new landscape. The good news is that there are good recipes in place, technologies do exist, processes do exist, and organizations have been able to deploy big data very successfully.

At Cloudera, we continue to develop the open-source, big data technology that’s proven and widely in use. And we offer the training, customization, and support you need to take full advantage of big data.

At Cloudera, we think it’s critical that the platform be 100 percent open source.

Build Your Big Data Knowledge and Hit the Ground Running

Cloudera is a pioneer in bringing open-source technology to the IT table, and most people running Hadoop today are running Cloudera’s distribution (CDH). This gives us a wealth of experience that we use to offer:

- **Cloudera Enterprise.** A simple subscription gives you operational support from Hadoop experts as well as rich management, monitoring, alerting, and integration capabilities. This is provided in the form of additional software that integrates our open-source technologies into your existing enterprise IT ecosystem.

- **Training for Hadoop.** Each year we train about 10,000 people through a number of courses with accompanying certifications. We teach people how to develop and how to administer the different open-source technologies we’ve developed.

- **Architectural services.** Our architectural services help your organization through the processes required to successfully run a big data environment.
Analyze Big Data in Real Time

We’ve recently ushered in a whole new layer of technologies that have really opened the door to real-time processing for big data. HBase* is a real-time database that gives you key value access at the millisecond level. We’ve seen customers develop some pretty sophisticated real-time applications such as recommendation engines and messaging to interact with these massive, complex volumes of data at real-time speeds. Flume enables you to stream data from a wide variety of data sources directly into HDFS* or HBase in real time. Flume scales to accommodate very large distributed infrastructure and collects any kind of data.

An Integrated System

CDH runs wherever the rest of your data infrastructure is located, whether that’s your data center, the cloud, or managed hosting. To enhance your CDH experience, we package a collection of Apache* technologies that work on top of Hadoop:

- **HBase** – Get precise atomic access to discrete data in real time.
- **Flume** – Move data in real time, streaming it into a cluster. Then stream that data into HBase for instant processing. Or you can stream it into Hadoop Distributed File System (HDFS) for collection and incremental processing.
- **Hive** – Use this SQL-like language to make Hadoop accessible to front-end business intelligence solutions. Hive has been certified with solutions from providers such as Tableau and MicroStrategy.
- **Pig Latin** – Process data procedurally with this extract, transform, load (ETL) dataflow-like language.
- **Oozie** – Utilize this server-based engine that provides fail-tolerant workflow and orchestration for the entire system.
- **ZooKeeper** – Employ this high-performance coordination system for distributed applications.
- **HUE** – Interact with the system using this rich graphical user interface with single sign-on and access to a Hive shell and file browser.
- **Mahout** – Get access to naive Bayes algorithms, k-means clustering, and random forests—all different types of machine learning and data mining capabilities that are typically used on big data.

Cloudera’s Integrated Hadoop* Stack
Why It’s Important for Big Data to Be Open

At Cloudera, we think it’s critical that the platform be 100 percent open source. Big data is changing the way customers think about how they manage data. Hadoop is opening the door to the idea that you actually can keep everything online and accessible. And you can process it in real time.

When we talk to customers, many of them are extremely wary of going with a solution that has any risk of vendor lock in. That’s why we spend a little more than half of our engineering dollars on things we give away. Cloudera’s contributions are now 60 to 70 percent of all vendor contributions to the Apache ecosystem.

From our perspective, big data solutions have got to be a case of “You own the platform. You own your data.”

We Couldn’t Do Any of This without Intel

Our CDH depends on Intel® processors in the CPU and at the networking layer. The open philosophy that Intel has been pursuing from a multi-core perspective lines up very well with the kind of distributed parallel processing that Hadoop envisions.

In addition, Hadoop requires a consistent network to operate as efficiently as it does. The high-performing, reliable, and consistent network backbone that Intel provides is crucial.

Profit from All Your Data

Cloudera’s focus is first to create an open-source, big data platform that’s ubiquitous. Then we add value by using our expertise to help you successfully put big data to work in your organization. When you do, you’ll be able to reveal new ways of doing business, improve interactions with customers, and create innovative products and solutions.

For more information about Cloudera’s Hadoop Distribution and other services, visit cloudera.com.

Share with Colleagues 😃linkedin share_email

This paper is for informational purposes only. THIS DOCUMENT IS PROVIDED "AS IS" WITH NO WARRANTIES WHATSOEVER, INCLUDING ANY WARRANTY OF MERCHANTABILITY, NONINFRINGEMENT, FITNESS FOR ANY PARTICULAR PURPOSE, OR ANY WARRANTY OTHERWISE ARISING OUT OF ANY PROPOSAL, SPECIFICATION, OR SAMPLE. Intel disclaims all liability, including liability for infringement of any property rights, relating to use of this information. No license, express or implied, by estoppel or otherwise, to any intellectual property rights is granted herein.

Copyright © 2012 Intel Corporation. All rights reserved. Intel, the Intel logo, Intel Sponsors of Tomorrow, and the Intel Sponsors of Tomorrow logo are trademarks of Intel Corporation in the U.S. and other countries.

*Other names and brands may be claimed as the property of others.