Dell In-Memory Appliance for Cloudera Enterprise 1.1

Hadoop Overview, Customer Evolution and Dell In-Memory Product Details

Author: Armando Acosta
Hadoop Product Manager/Subject Matter Expert
Armando_Acosta@Dell.com/ Twitter- @Armando75

Overview

This paper is intended to provide an overview of the Dell In-Memory Appliance for Cloudera Enterprise 1.1 with Spark support. This paper is intended for Hadoop users and customers looking for a better understanding of the Dell In-Memory Appliance built on Cloudera Enterprise with Spark including use cases.

The paper will cover a basic overview of Hadoop, the customer evolution utilizing Hadoop, and the Dell in-Memory Appliance for Cloudera Enterprise. This paper is not a solution sizing document or a deployment guide for the Dell In-Memory Appliance for Cloudera Enterprise.

Customers with interest in deploying a proof of concept (POC) can utilize the Dell Solution Centers. The Dell Solution Centers are a global network of connected labs that enable Dell customers to architect, validate and build solutions. Customers may engage with their account team and have them submit a request to take advantage of the free services.

Executive Summary

Data is the new currency and competitive differentiator, data is being created and consumed at rates never before seen. This is not unique to a single industry; it will affect all vertical markets. Organizations are struggling to ingest, store, analyze, and build insights from all this data. As more connected devices and machines with embedded sensors proliferate throughout the world, this will create even greater challenges for customers. Dell, together with Cloudera and Intel, want to help customers solve this problem with a turnkey, purpose built in-memory advanced analytics data platform. The Dell In-Memory Appliance for Cloudera Enterprise 1.1 represents a unique collaboration of partners within the big data ecosystem, delivering both the platform and the software to help enterprises capitalize on high-performance data analysis by leveraging the Cloudera Enterprise Data Hub's in-memory features, Spark, Impala, and Cloudera Search, for interactive analytics and multiple types of workloads.

Customers continue to struggle with the deployment, configuration, tuning, and optimizing of Hadoop distributions and clusters. Customers want faster deployment of the solution that will then allow them to focus on the analysis of the data. Customers want big data solutions that easily integrate and can quickly start delivering value by cutting deployments from months to weeks or even days.

Additionally, data analysis is a complex process made up of many steps and various tools that allow users to do different types of analysis. In order to accomplish work, businesses build a workflow comprised of many tools, including different steps for different types of analysis. The different tools complicate the workflow because the users must provide all the translation to give context to each different tool. This affects the productivity and the speed to deliver the results a business needs in order to make proactive decisions.

The Dell In-Memory Appliance for Cloudera Enterprise 1.1 is a preconfigured hardware and software stack that takes the time and effort out of deployment, configuration, tuning, and optimizing of a Hadoop distribution and cluster for streaming workloads. Cloudera Enterprise support of Spark, Impala, and Cloudera Search helps simplify customer environments by giving them one single tool for data processing and interactive analysis. Dell Services quickly integrates the appliance into a customer's environment speeding deployment thus reducing the time it takes compared to building a bare metal cluster from scratch where deployment includes loading O/S, BIOS, FW, and Hadoop distribution plus network switch configuration.

Apache Hadoop

Hadoop is an Apache open-source project being built and used by a global community of contributors who use the Java programming language. Hadoop was originally developed by Yahoo!® and provided to the open-source community to help solve the need for a scalable, highly distributed file system (HDFS) with a massively parallel computing framework, MapReduce. Both HDFS and MapReduce implementations were inspired by papers written and published by Google®. The solution allowed both companies to ingest massive amounts of structured, semi-structured, and unstructured data while being able to process the data in an efficient manner. Other contributors and users include

Facebook®, LinkedIn®, eHarmony®, Intel® and eBay®. Hadoop's architecture is based on the ability to scale in a nearly linear capacity. Harnessing the power of this tool enables customers who would have previously had difficulty sorting through their complex data, to now deliver value faster, provide deeper insights, and even develop new business models based on the speed and flexibility these analytics provide. Another significant benefit is that Hadoop allows users to store all types of data on a single platform rather than using multiple tools that do not scale and are not cost-effective.

Apache Hadoop is built from a set core of components that make up a set of tools that allow customers to build solutions for their use cases. Most Hadoop users utilize a combination of these tools, with packaged software, based on their set of needs.

Table 1. Hadoop Core Components and Ecosystem Tool Descriptions

Component	Description
Apache Hadoop (HDFS)	Framework for reliable, scalable storage of large data sets for distributed computing
Apache MapReduce v2	Programming model for distributed data processing on large cluster of compute nodes
Apache HBase™	Distributed scalable Big Data store for random real-time read/write access
Apache Hive™	Data warehouse system for query using SQL-like language
Apache Pig™	Platform for analyzing large data sets consisting of high-level language for expressing data analysis programs
Apache ZooKeeper™	Centralized service for highly reliable distributed coordination
Apache Sqoop™	Tool for bulk data transfer between Apache Hadoop and relational databases
Apache Flume™	Distributed reliable and available service for collecting, aggregating and transferring large amounts of log data
Apache Mahout™	Library of machine learning for Apache Hadoop
Apache Oozie	Workflow system scheduler to manage Apache Hadoop jobs
Apache Lucene™, Apache Solr™,	Enterprise search platform
Apache Spark™	Real-time processing engine for large-scale data processing ideal for cluster computing
Apache Shark™	Distributed SQL query engine
Apache Kafka™	High-throughput distributed messaging system

When MapReduce, or Apache Spark, is paired with HDFS, the result provides a high-performance solution running across a cluster of commodity x86 servers; this helps lower costs, a key factor in the popularity of Hadoop. One of the keys to Hadoop performance is the lack of data motion where compute tasks are moved to the servers on which the data resides. MapReduce tasks, or Spark jobs, can be scheduled on the same physical nodes on which data are resident in HDFS. This design

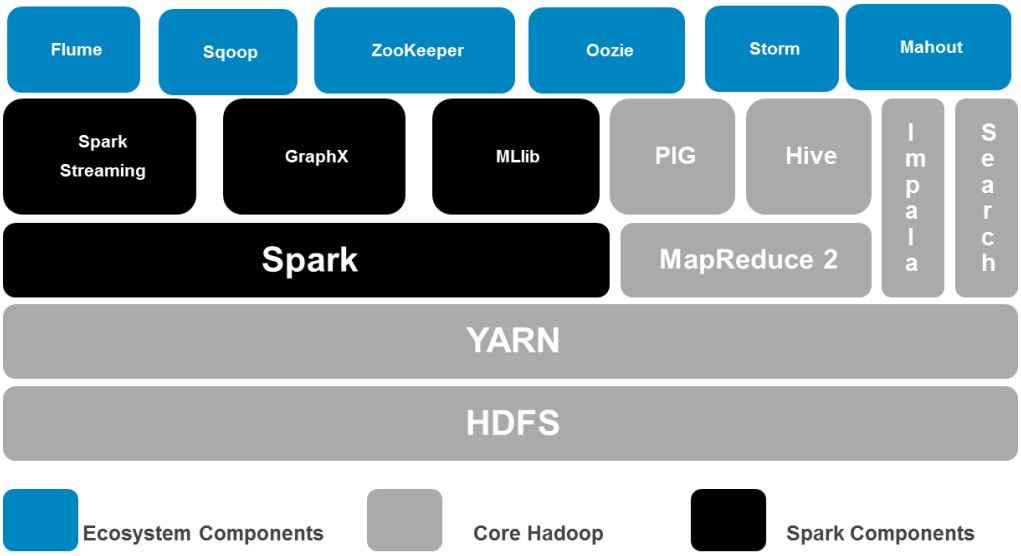
significantly reduces the network traffic and keeps most of the I/O on the local disk or on a server within the same server rack.

With the introduction of YARN (**Y**et **A**nother **R**esource **N**egotiator), Spark jobs can be spun up to be processed in-memory. Spark utilizes HDFS enabling customers to utilize the existing data within HDFS without movement prior to data analysis.

The Apache Spark project is a fast and general purpose engine for large-scale data processing. Spark allows users to do in-memory computing resulting in programs running up to 100x faster than Hadoop MapReduce in-memory, or 10x faster on disk. Spark powers a stack of high-level tools including Spark SQL, MLlib for machine learning, GraphX for graph analysis, and Spark Streaming for stream processing. You can combine these frameworks seamlessly in the same application.

Spark can also be deployed via Apache Mesos or in stand-alone mode. The Dell In-Memory Appliance for Cloudera Enterprise 1.1 with Spark support provides simplified scalability for building data pipelines that include both HDFS and Spark processing. This benefit is not possible with stand-alone mode.

Figure 1: Hadoop Core Components, Ecosystem Components, and Spark Components



Customers Evolving with Hadoop

Hadoop is no longer a technology that exclusively belongs in Web 2.0; it has started to shift into more traditional IT environments. As this shift occurs, customer needs are changing based on use cases and the need to analyze data in as fast and in as efficient a manner, within a short window of time, and on a continuous basis.

It's very important to take a historical perspective to understand the development of Hadoop and the customer maturity with the technology.

Customer Needs Have Changed Over Time

2007-2010: Customers start hearing about Hadoop- (Innovators)

- What is it?
- How do you use it?
- How do you build it?
- Can it make difference for my organization?
- How much data and what type of data would I put in Hadoop?

2010 -2013: Customers start adopting Hadoop- (Early Adopters)

- Customers start executing proofs of concept.
- Customers now consider data types and sources. Ingest/Export operations.
- Customers determine the type of analytics tools they might use with Hadoop.
- There is a lack of Hadoop/Ecosystem expertise and skill set.

2014 ~2015: Company-wide initiatives, data is the new currency- (Mainstream)

- Customers need a simpler way to deploy a Hadoop solution.
- Customers need to stop spending time on deployment, configuration, tuning, and optimization.
- Customers are developing Hadoop/ecosystem expertise.
- Hadoop teams are working with data scientists/SW developers to simplify analysis and reduce workflow steps.

2016 ~beyond: Data continues to grow; batch processing is not enough- (Core Business Strategy)

- Customers will still be struggling to ingest, store, analyze, and build insights from all this data.
- As more connected devices and machines with embedded sensors proliferate throughout the world, this will create even greater challenges for customers.
- The Internet of Things will drive a need for fast processing of data and analytics.
- Data analysis will become a core business function and key piece of the decision process.

Hadoop with Cloudera Enterprise plus Spark will be able to fill those many customer needs today and in the future. Spark utilizes in-memory computing to deliver high performance data processing with analysis. Within the Spark computing framework, it is also tooled with analytics packages for interactive query, iterative processing, graph analysis and streaming data. Spark will allow customers to use one tool for all the different types of analysis that are required of their data.

Dell In-Memory Appliance for Cloudera Enterprise

The Dell In-Memory Appliance for Cloudera Enterprise 1.1 is a preconfigured hardware and software stack that takes the time and effort out of deployment, configuration, tuning, and optimization of a Hadoop distribution and cluster for streaming workloads. Cloudera Enterprise support of Spark helps simplify customer environments by giving them one single tool for data processing and interactive

analysis. Dell Services quickly integrates the appliance into a customer's environment speeding deployment.

Dell Engineered Systems have built, tested, validated, tuned, and optimized a solution in an appliance delivery model. The Dell In-Memory Appliance for Cloudera Enterprise 1.1 is a fully configured, racked & stacked, enterprise ready production cluster that includes the networking running Cloudera Enterprise. The in-memory appliance is delivered as one solution to the customer site, then quickly integrated by connecting ToR switching and assigning IP addresses, to finalize the Hadoop distribution configuration.

The Hadoop data node configurations have been optimized to run streaming workloads that require more CPU processing and additional memory capacity that then allow customers to analyze larger data sets.

Cloudera Enterprise, Impala and Spark are pre-configured and ready to use after integration services. Cloudera Manager is fully functional and configured reducing the time to start processing and analyzing the data. All Hadoop cluster services will be assigned running on the pre-configured infrastructure nodes, customer will not have to spend valuable time on this set-up.

The Dell In-Memory Appliance for Cloudera Enterprise 1.1 takes the time and effort out of deployment, configuration, tuning, and optimizing of a Hadoop distribution and cluster for streaming workloads. The solution is built to be highly redundant and tuned to run streaming workloads efficiently.

Solution Components

The Dell In-Memory Appliance, powered by Cloudera Enterprise, solution architecture includes Dell servers with internal storage, software, services, and networking infrastructure.

The appliance comes in three configurations: Starter Configuration, 8 nodes, Mid-Size Configuration. 16 nodes, and Small Enterprise Configuration, 24 nodes. For additional capacity, the Expansion Unit of 4 data nodes can be added. This allows customers to build their big data solutions based on PODs for multiple streaming workloads or for a proof of concept. The solution can scale up to 48 Nodes.

The pre-configured solution includes Dell PowerEdge R730/R730XD servers, Dell Network switch technology, Dell Services for on-site customer integration, and Cloudera Enterprise.

Dell PowerEdge R730/R730XD

The PowerEdge R730 and R730xd servers are Dell's 13G PowerEdge dual socket 2U rack servers. They are designed to deliver the most competitive feature set, best performance and best value in this generation, Dell offers a large storage footprint, best-in-class I/O capabilities and more advanced management features. The PowerEdge R730 and R730xd are technically similar except the R730xd has a backplane that can accommodate more drives (up to 24).

Dell Networking

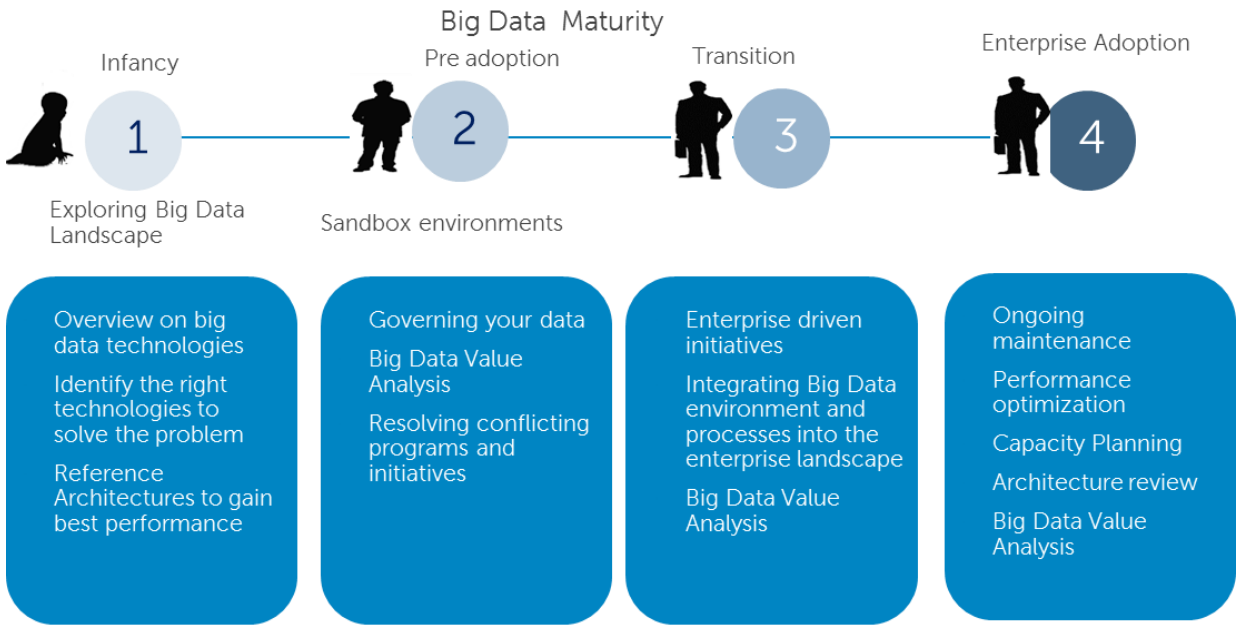
The Dell Networking S-Series S4048 is an ultra-low-latency 10/40GbE top-of-rack (ToR) switch purpose-built for applications in high-performance data centers and computing environments. Leveraging a non-blocking, cut-through switching architecture, the S4048 delivers line-rate L2 and L3 forwarding capacity with ultra-low latency to maximize network performance. The compact S4048 design provides 48 dual-speed 1/10GbE (SFP+) ports as well as four 40GbE QSFP + uplinks to conserve valuable rack space and simplify the migration to 40Gbps in the data center core.

Dell Services and Support

Dell Services provide integration services on the customer's site to quickly integrate the Dell In-Memory Appliance for Cloudera Enterprise.

Optional, Hadoop Administration Services (HAS) are available and are designed for customers who are embarking on the big data journey or are already deriving value from big data. This service is designed to support the customer from the inception to steady state.

Figure 2: Hadoop Administration Services (HAS) for Big Data Maturity



Big Data Maturity - Ability to integrate, manage and leverage data to drive business value

Dell Support provides the highest level of support to help keep your solution up and running.

ProSupport Plus

- Technical Account Manager
- Dell supports Hardware and RHEL O/S
- Collaborative Assistance with Cloudera for Hadoop distribution

Cloudera Enterprise

Cloudera Enterprise includes CDH, the world's most popular open source Hadoop-based platform, as well as advanced system management and data management tools plus dedicated support and community advocacy from our world-class team of Hadoop developers and experts.

CDH is the world's most complete, tested, and popular distribution of Apache Hadoop and related projects. CDH is 100% Apache-licensed open source and is the only Hadoop solution to offer unified batch processing, interactive SQL, and interactive search, and role-based access controls. CDH is thoroughly tested and certified to integrate with a wide range of operating systems and hardware, databases and data warehouses, and business intelligence and extract, transform and load (ETL) systems.

Cloudera Manager is designed to make the administration of CDH simple and straightforward, at any scale. With Cloudera Manager, you can easily deploy and centrally operate the complete Hadoop stack. The application automates the installation process, reducing deployment time from weeks to minutes; gives you a cluster-wide, real-time view of nodes and services running; provides a single, central console to enact configuration changes across your cluster; and incorporates a full range of reporting and diagnostic tools to help you optimize performance and utilization.

Cloudera Impala is an open source Massively Parallel Processing (MPP) query engine that runs natively in Apache™ Hadoop®. The Apache-licensed Impala project brings scalable parallel database technology to Hadoop, enabling users to issue low-latency SQL queries to data stored in HDFS and Apache HBase™ without requiring data movement or transformation. Impala is integrated from the ground up as part of the Hadoop ecosystem and leverages the same flexible file and data formats, metadata, security and resource management frameworks used by MapReduce, Apache Hive™, Apache Pig™ and other components of the Hadoop stack.

Cloudera Search brings full-text, interactive search and scalable, flexible indexing to CDH and your enterprise data hub. Powered by Apache Hadoop and Apache Solr, the enterprise standard for open source search, Cloudera Search brings scale and reliability for a new generation of integrated, multi-workload search. Through its unique integrations with CDH, Cloudera Search gains the same fault tolerance, scale, visibility, security, and flexibility provided to other enterprise data hub workloads. Users are free to examine vast amounts of content and attributes to discover the hidden correlations between the data points in near real-time. Faceted navigation surfaces – content attributes to assist and guide users as they explore data – allows users to discover the "shape of data" quickly and easily and expedite data modeling and result exploration.

Dell In-Memory Appliance 1.1 Configuration Detail:

Starter Configuration- is ideal for exploring and testing new Spark or streaming workloads in a proof of concept environment or small production environment

- 8 Node Cluster
- PowerEdgeR730- 4 Infrastructure Nodes, PowerEdgeR730XD- 4 Data Nodes
- Cloudera Enterprise
- 2 x Dell Networking - S4048P
- 1 x Dell Networking - S3810

- 1 x Dell Rack 42U
- Memory ~1.5TB (Raw)
- ProSupport Plus

Mid-Size Configuration- This robust and scalable configuration is designed for a production environment

- 16 Node Cluster
- PowerEdgeR730- 4 Infrastructure Nodes, PowerEdgeR730XD- 12 Data Nodes
- Cloudera Enterprise
- 2+ x Dell Networking - S4048P
- 1 x Dell Networking - S3810
- 2 x Dell Rack 42U
- Memory ~4.5TB (Raw)
- ProSupport Plus

Small Enterprise Configuration- This robust and scalable configuration is designed for a production environment

- 24 Node Cluster
- PowerEdgeR730- 4 Infrastructure Nodes, PowerEdgeR730XD- 20 Data Nodes
- Cloudera Enterprise
- 2+ x Dell Networking - S4048P
- 1 x Dell Networking - S3810
- 3+ x Dell Rack 42U
- Memory ~7.5TB (disk raw space)
- ProSupport Plus

Summary

With the Dell In-Memory Appliance for Cloudera Enterprise 1.1, Dell, Cloudera and Intel deliver the platform and the software that can help enterprises capitalize on high-performance data analysis. The capabilities and opportunities available to organizations leveraging the appliance capitalize on the inclusion of Cloudera Enterprise with the Data Hub in-memory features, Spark, Impala, and Cloudera Search, to enable interactive analytics on multiple types of workloads.

To learn more, or for additional information, please visit Dell.com/Hadoop or contact: Hadoop@Dell.com.

SOURCES:

RESILIENT DISTRIBUTED DATASETS: A FAULT-TOLERANT ABSTRACTION FOR IN-MEMORY CLUSTER COMPUTING; MATEI ZAHARIA, MOSHARAF CHOWDHURY, TATHAGATA DAS, ANKUR DAVE, JUSTIN MA, MURPHY MCCAULEY, MICHAEL J. FRANKLIN, SCOTT SHENKER, ION STOICA UNIVERSITY OF CALIFORNIA, BERKELEY

FAST AND INTERACTIVE ANALYTICS OVER HADOOP DATA WITH SPARK; TATHAGATA DAS, ANKUR DAVE, JUSTIN MA, MURPHY MCCAULEY, MICHAEL J. FRANKLIN, SCOTT SHENKER, ION STOICA

DATABRICKS WEBSITE- www.databricks.com/spark

APACHE SPARK WEBSITE- www.spark.apache.org

CLOUDERA SPARK WEBCAST- SPARK, THE NEXT GENERATION OF MAPREDUCE