HITACHI
Inspire the Next

How Hitachi Data Systems Hadoop Solution

WebTech Q&A Session, June 12, 2013

1. **What are the most common use cases for Hadoop?**
   Hadoop has many use cases, but the 3 most common are the following:
   Improved Customer Analytics – enhancing the understanding of customers. This is feasible by integrating behaviors that are captured via traditional transactions, irrespective of volume and the rate in which they flow, with spontaneous interactions through social media, irrespective of the structure.

   Improved Operations Analytics – ability to proactively identify and respond to critical events. This is feasible as more and more machine-generated data becomes available that is well suited to being handled by the technologies and platform we discussed. Hitachi Consulting has domain experts who understand the nuances of data and processes when defining and implementing a solution.

   Enhanced Data Warehouse environments – extending current, and introducing new, capabilities. This is feasible because Hadoop-based platforms let us handle entire datasets instead of subsets. The platform's scalability and flexibility enables us to overcome the structural limitations traditional data warehouses impose: structure-on-load versus structure-on-read paradigm. Plus, Hadoop-based solutions generally help lower total cost of ownership as well.

2. **Is Cloudera's software open source?**
   Cloudera's platform is both 100% open source and Apache licensed. In addition to its open-source platform, Cloudera also offers proprietary add-on tools that make the Apache Hadoop platform simpler to manage and use.

3. **What's the minimum cluster size to get started with Hadoop?**
   We recommend at least 3 data nodes to get started with Hadoop. This should provide adequate data processing capability and data protection.

   Apache Hadoop is capable of running with as little as a single node. Just to get started, 1 to 5 nodes would be sufficient to get familiar with the platform. Running 4 to 10 nodes is generally enough to prove the platform's capabilities and to begin seeing value. The typical initial cluster size for a trial or Proof-of-Concept (PoC) is 10 to 20 nodes.

4. **You talked about an adoption strategy for big data technologies that protects some of our business intelligence (BI) investments. Can you elaborate?**
   This relates to the use case we talked about: enhancing data warehouse environments by extending current, and introducing new, capabilities in a gradual manner. For example, consider a scenario where you already have reasonably well-functioning customer-analytics assets, such as customer-

relationship datamart models, as well as customer-analytics cubes and reports based on data from purpose-built transactional systems. You could then integrate customer interactions using social media data, which is primarily unstructured, through additional components by leveraging the big data capabilities we talked about. These additional components get data from social media sources and map customer behaviors to pre-existing measures. With reasonably minimal updates to your existing assets – such as data models, ETL components and reports – you can enhance your understanding of your customers.

5. **What are the high availability options for the Name node?**

   Cloudera's Distribution including Apache Hadoop (CDH) includes a highly available (HA) Name Node configuration option that provides full redundancy for the Name Node.

6. **What Hadoop distributions are supported by Hitachi?**
   Hitachi reference architecture for Hadoop is agnostic about Hadoop distribution and can deploy any Apache Hadoop-based distribution. Several vendors – such as Cloudera, HortonWorks, MapR, and Intel – take subsets of the Apache Hadoop family of technologies and bundle them with patches and additional software. Hitachi reference architecture is currently being certified with Cloudera, but other distributions will also work

7. **Is Cloudera using Hadoop Distributed File System (HDFS)? How does Cloudera provide file consistency in case of cluster node failure? How long does it take the node to rejoin the cluster?**

   HDFS is the core of Apache Hadoop, and CDH is built around it. HDFS prevents data loss in the event of node failure by replicating data blocks across nodes and racks. When a node fails, the data on that node is still available from nodes with replicas, and HDFS works to restore the minimum replica count. Because data blocks in HDFS are immutable, when a failed node is brought back into service, there's no barrier to keep it from rejoining the cluster.

8. **What BI reporting tool does this solution use?**

   CDH supports all the usual interfaces that BI tools use to connect with data sources, such as Java Database Connectivity (JDBC) and Open Database Connectivity (ODBC). Name any BI or analytics tool, and odds are good that Cloudera is working with them and that they support CDH. In addition to the Hive SQL engine that is common to most Hadoop distributions, Cloudera has also introduced the Impala project, which provides real-time SQL query capabilities over HDFS data. Cloudera Impala 1.0 is generally available and is also Apache-licensed open source.

9. **What is the relationship between Cloudera and HDS?**
   HDS is a Cloudera partner. Hitachi reference architecture is being certified with Cloudera distribution.

10. **How many server spares can you have in a HDFS cluster? Can 2 or more servers fail without causing downtime or loss of data?**
    HDFS is a distributed file system where files are replicated across multiple nodes to avoid data loss or downtime. You can set the replication of data to 3 and the data files will be replicated across 3 random nodes, typically 2 copies in the servers within the same rack and the 3rd copy in a different rack. For high data availability, you can increase the data replication factor to 4 or more. All the

servers in a Hadoop cluster are active. If one node fails, the task will be delegated to the node that has the 2nd or 3rd copy of the data. So you do not have spare servers, per se, in a HDFS cluster.

11. **Because of the Hadoop cluster "data locality" and "shared nothing" principle, using a SAN or NAS in a Hadoop environment was NOT recommended. For Hadoop deployments using a SAN or NAS, the extra network communication overhead can cause performance bottlenecks, especially for larger clusters. Has Hadoop broken the rule to accommodate with SAN or NAS?**

    The massive performance gains often seen from porting applications to Hadoop are often due to the data locality concept, which significantly reduces the need to move data across the network. Running HDFS across a SAN or NAS environment negates data locality and hence a large portion of the performance gains.

12. **In the reference architecture, each node has 12 disks. Those disks are DAS (directly attached storage) or SAN/NAS?**
    These are direct-attached disks. High-performance Hadoop architecture requires keeping the data close to the CPU to reduce latency. As a result, DAS is used.