Building & Optimizing Enterprise-class Hadoop with Open Architectures

Prem Jain
NetApp
Introduction to Hadoop

- Comes from Internet companies
- Emerging big data storage and analytics platform
- HDFS and MapReduce and key components
- Apache foundation project
- Key players are Cloudera, Hortonworks, MapR, IBM, Pivotal etc.
- Experiencing explosive growth
- Growing adoption into enterprise
Traditional Hadoop Architecture

- **NameNode**
- **Secondary NameNode**
- **Job Tracker**

**Datanodes**

- **Fat Servers**
- **Network Congestion**

**Server Replication R=3 (typical)**
Challenges with Hadoop in Enterprise

<table>
<thead>
<tr>
<th>Operational</th>
<th>Efficiency</th>
<th>Manageability</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Limited flexibility of scaling</td>
<td>• Three copies of data</td>
<td>• Requires Skilled resources</td>
</tr>
<tr>
<td>• Jobs get interrupted during failures</td>
<td>• Network congestion</td>
<td>• Less predictable SLA’s</td>
</tr>
<tr>
<td>• Flexibility of changing architectures</td>
<td>• Low storage efficiency</td>
<td>• Data center footprint</td>
</tr>
</tbody>
</table>
Hadoop in the Enterprise

1,000 Commodity Servers

1,000-Node Cluster of Commodity Servers

Science Project
- Skill set
- Hadoop data management
- Data ingest
- Cluster management

Cluster Inefficiency
- Scale primary motivation
- Single knob: “multiple replicas”
- Ingest bottlenecks
- Inefficient network usage

TCO Issue
- Economies of scale?
- Scale opex with cluster?
- Opportunity cost?
NetApp Hadoop Architecture

NameNode

Secondary NameNode

Job Tracker

Datanodes

Storage Array

Storage Array
NetApp Hadoop Architecture

- NameNode
- Secondary NameNode
- Job Tracker
- Datanodes
- Storage Array
- Server Replication 2 or None
- Compute
- Storage
Flexible and open architecture

- No vendor lock in
- Optimized infrastructure investments
- Flexibility to changes
- Reduce risk in an open ecosystem
Key advantages

- Independent scaling of compute & storage
- RAID based storage reliability
- Lesser replicas of data
- Guaranteed performance even during failures
- Lower management overhead
- Enterprise level infrastructure
- SLA based management
Enterprise Hadoop Platforms

- Apache Hadoop
  - HDFS
  - 10 GbE
  - UCS C220 M3 Servers
  - Namenode
  - Secondary Namenode
  - Datanodes/Tasktrackers

- FAS® Series
  - Data ONTAP
  - FAS2240
  - Metadata Store
  - Highly available
  - E-Series
  - E5460
  - Data Stores
  - 4 separate, shared-nothing platforms per chassis
  - E5460
  - 6 GB SAS Direct Connect

- Expansion
  - Cisco Fabric Interconnect
  - Metadata Store
  - Name Node and secondary Name Node
  - Data Nodes
  - Data Store
Hadoop Open Architectures
Validated for Cloudera and Hortonworks

- Converged big data platform from NetApp and Cisco for Hadoop
- Enterprise class Hadoop: Innovative storage, servers, networking validated with leading Hadoop distributions
- Faster time to value: pre-validated configuration accelerates deployment
- High Availability: Less downtime, higher serviceability to meet tight SLAs around data applications and processes
- Flexible Scaling: Independently scale servers and storage. Modular design for scaling as data needs grow.
Thanks