Data Services for Hybrid Cloud

Ameya Prakash Usgaonkar
Principal Engineer, NetApp
May 2017
Agenda

1) Introduction to Hybrid Cloud
2) Hybrid Cloud & Data Fabric
3) Hybrid Cloud Workloads & Challenges
4) Data Service – Hybrid Cloud’s Next Generation Data Management
5) Data Services Platform Architecture
6) Case Study – Big Data Service
7) Summary
Introduction to Hybrid Cloud

Fundamental Trend for IT

RESPONSIVENESS

Agile Application Delivery

Cloud Service Providers

Hyperscale Cloud Providers

Speed Innovation

Cost reduction

Turn Capex to Opex

Private

Public

Elastic Management

Pay as you go

Private Cloud

CHOICE

CONTROL

© 2017 NetApp, Inc. All rights reserved. NetApp Confidential
Data and Hybrid Cloud

Unique Requirements

Cloud Service Providers

Hyperscale Cloud Providers

Private Cloud

DATA

Security

Access

Protection

Availability

Compliance

Governance
Data and the Hybrid Cloud

The need for a data fabric

Cloud Service Providers

Hyperscale Cloud Providers

Private Cloud
What Does a Data Fabric Do?

Consistent uniform data management

Realize the Full Potential of Hybrid Cloud
Hybrid Cloud Workloads and Challenges

Isolated, Incompatible data silos

- Sprawls of Copies
- Redundant n/w transfers
- Data Tracking
- Co-location Investment
- Data Transformation
- Corporate IT Involvement
Data Service – Next Generation Data Management for Hybrid Cloud Workloads

- Storage endpoint agnostic discrete software catering to specific workloads in hybrid cloud
  - Built on fundamental storage APIs

- Extensive use of metadata to perform on-the-fly transformation of data
  - Richness and proximity of metadata determine value of service and performance respectively

- Suited for DevOps style workflow without IT involvement
  - Ease of deployment, Faster innovation
  - Integration with cloud services for hosting in hybrid cloud

- Collection of services constitutes “Data Services Platform” for Data Fabric
Data Services Platform Architecture

Data Service-1  Data Service-2  Data Service-3  Data Service-4  Data Service-5  Data Service-6

Metadata Store

Metadata APIs
(POSIX/Block/Object/KV)

Storage APIs
(Snapshot, Clone etc.)
Case Study – Big Data Service

- **Problem Statement**
  - Analytics in hybrid cloud
    - Enable Big Data Analytics of on-premise data *directly* in cloud without E-T-L

- **Challenges**
  - Avoid complex data management (data tracking, security etc.)
  - Increase responsiveness – no E-T-L phase
  - Use on-premise data management AND elastic resource management of cloud (best of both worlds)

- **Solution**
  - Provide a cost-effective and performant solution of running Big Data Analytics in cloud while continuing to manage the data in the private data center
Technical Requirements

- No E-T-L (No data copy)
- Horizontal Scale Out
- Heterogeneous (storage endpoint agnostic)
- WAN friendly data transfer (compression and caching)
- Corporate firewall friendly data transfer protocol
- Software-defined DevOps style workflow
- Ease of deployment
Experimental Setup

- **Hadoop Cluster (version 2.7.2)**
  - 5 nodes (1 master, 4 slaves)
  - EC2 Node: 16 vCPUs, 64GB RAM, 2 Gbps network throughput, 300GB EBS

- **Metadata Software**
  - Server running Ubuntu Linux, 2 vCPUs, 8GB RAM connected to NetApp FAS
  - Metadata setup time is 40 minutes
  - Secure on-demand data transfer over firewall friendly protocol/s (HTTPS, SSL, SFTP etc.)

- **E-T-L Details**
  - Two files, each of 100 GB, generated by Hadoop TeraGen
  - "Rsync" transferred dataset from NetApp lab to EBS in US-East-1 in N. Virginia @20MB/s and 75ms latency in 200 minutes
  - Time to ingest 200 GB from EBS to HDFS is 27 minutes

- **Capture MR time, CPU time, Total time and Memory consumption**
  - Applications: TeraSort and WordCount
Big Data Service – Experimental Results (TeraSort)

Job Summary Time (minutes) for Terasort (Dataset: 200GB)

<table>
<thead>
<tr>
<th>Task</th>
<th>Native HDFS</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset transfer time to AWS EBS / Metadata Setup Time</td>
<td>200</td>
<td>40</td>
</tr>
<tr>
<td>Data ingest time to HDFS</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Map Time</td>
<td>13</td>
<td>46</td>
</tr>
<tr>
<td>Reduce Time</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>Tera Sort (MR) Time</td>
<td>27</td>
<td>69</td>
</tr>
<tr>
<td>Overall Time (1+2+3+4)</td>
<td>254</td>
<td>109</td>
</tr>
</tbody>
</table>

Resource Consumption

<table>
<thead>
<tr>
<th>Resource</th>
<th>Native HDFS</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Time (seconds)</td>
<td>18402</td>
<td>15410</td>
</tr>
<tr>
<td>Physical Memory (GB)</td>
<td>1501</td>
<td>1479</td>
</tr>
<tr>
<td>Virtual Memory (GB)</td>
<td>4147</td>
<td>4146</td>
</tr>
<tr>
<td>Total Heap (GB)</td>
<td>1702</td>
<td>1625</td>
</tr>
</tbody>
</table>
Job Summary Time (minutes) for WordCount (Dataset: 200GB)

- Native HDFS
- Prototype

<table>
<thead>
<tr>
<th>Task</th>
<th>Native HDFS</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Dataset transfer time to AWS EBS / Metadata Setup Time</td>
<td>200</td>
<td>259</td>
</tr>
<tr>
<td>(2) Data ingest time to HDFS</td>
<td>40</td>
<td>107</td>
</tr>
<tr>
<td>(3) Map Time</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>(4) Reduce Time</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>Word Count (MR) Time (3+4)</td>
<td>49</td>
<td>67</td>
</tr>
<tr>
<td>Overall Time (1+2+3+4)</td>
<td>20609</td>
<td>18733</td>
</tr>
</tbody>
</table>

Resource Consumption

- Native HDFS
- Prototype

<table>
<thead>
<tr>
<th>Resource</th>
<th>Native HDFS</th>
<th>Prototype</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU Time (seconds)</td>
<td>1884</td>
<td>1973</td>
</tr>
<tr>
<td>Physical Memory (GB)</td>
<td>1761</td>
<td>1878</td>
</tr>
<tr>
<td>Virtual Memory (GB)</td>
<td>4147</td>
<td>4146</td>
</tr>
<tr>
<td>Total Heap (GB)</td>
<td>1761</td>
<td>1878</td>
</tr>
</tbody>
</table>
Representative Workloads in Hybrid Cloud

Combine performance and reliability of dedicated servers with elasticity of cloud hosting solutions

Big Data
Geo Caching
Test and Dev
DevOps
Genomics
Data Tiers

Elastic Resource Management
Summary

✓ Hybrid cloud model requires software-defined data delivery and management without corporate IT involvement

✓ Data Service is next generation storage endpoint agnostic, workload centric data management software deployed using DevOps style workflows in hybrid cloud

✓ In this presentation, we presented Big Data Service for hybrid cloud without complicating data management for IT

✓ NetApp Data Fabric enables enterprises to build a foundation for hybrid cloud today, and then use it to connect to variety of services in public cloud
Thank You.