NetApp

SPARK SUMMIT 2017

ANALYZING IOT DATA IN APACHE SPARK ACROSS DATA CENTERS AND CLOUD

By

Karthikeyan Nagalingam

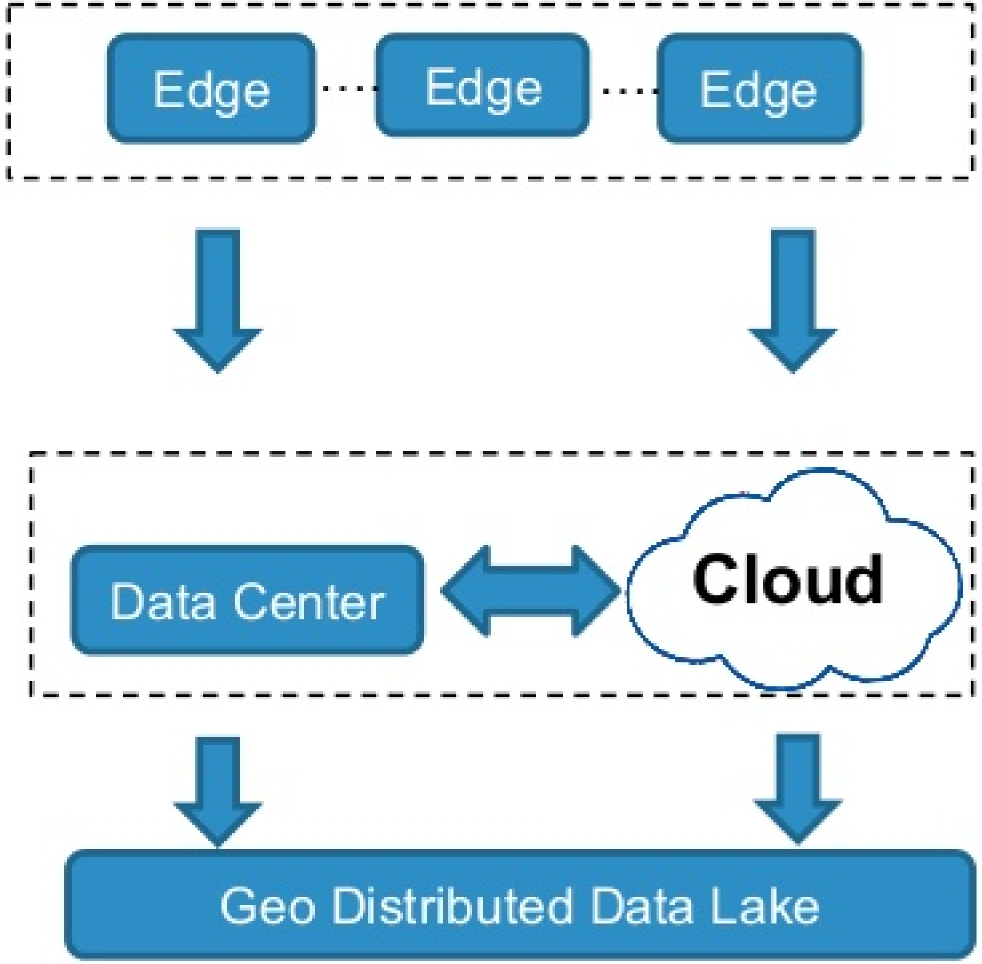
Nilesh Bagad

Agenda

- IoT Data Management Challenges
- NetApp Data Fabric Architecture for Big Data
- IoT Customer Use Cases on Data Fabric
- Q&A

IoT Data Management Challenges

IoT Data Flow



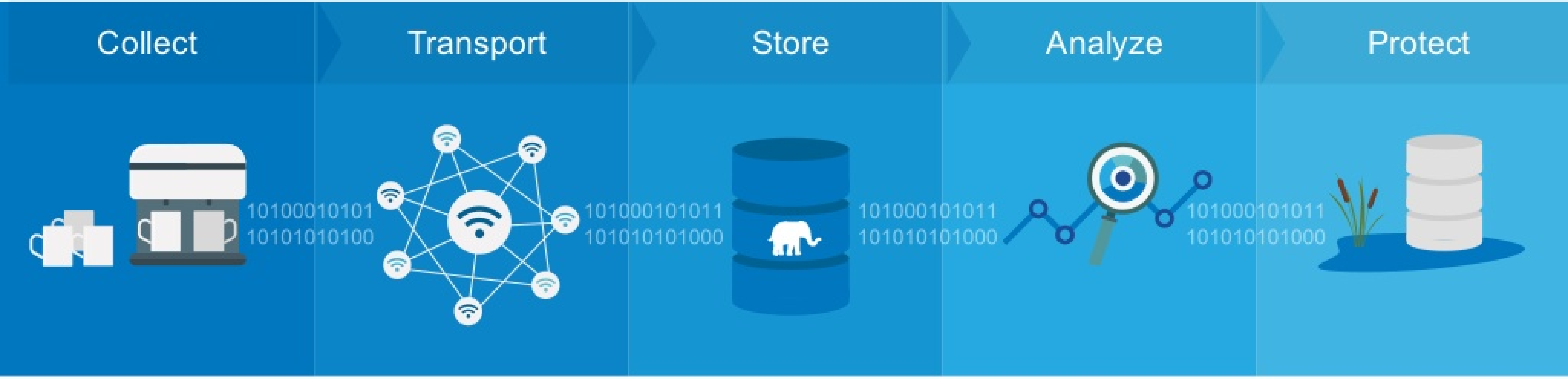
EDGE

1. Data is created
2. Data is analyzed in realtime
3. Data is aggregated and sent to Core

CORE

1. Data is stored
2. Data is analyzed
3. Data is protected

IoT Data Management Challenges



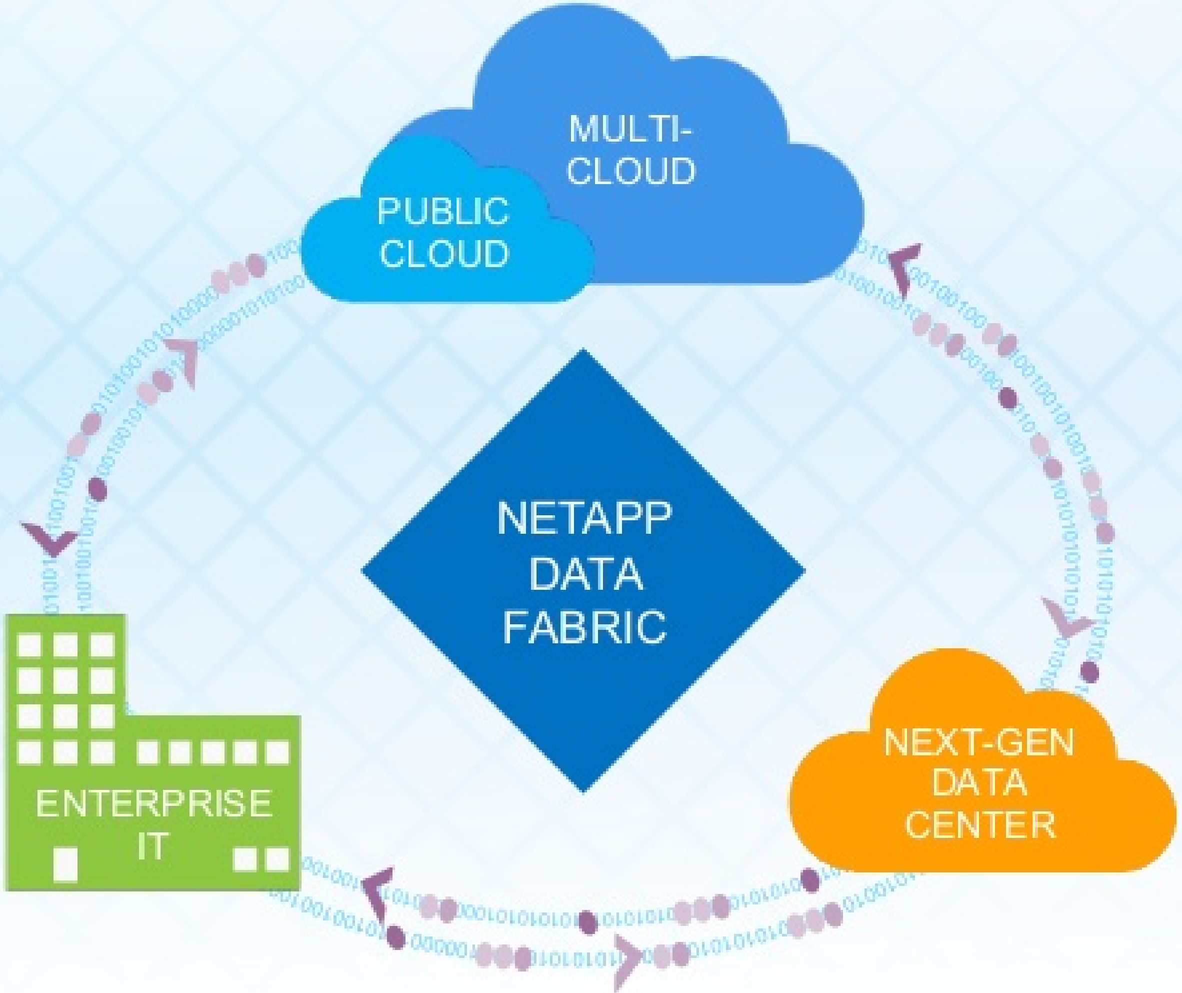
- **Flexibility and Agility**
- **Cost**
- **Data Protection**

NetApp Data Fabric Architecture for Big Data

SPARK SUMMIT 2017

The NetApp Data Fabric

Helping customers unleash data to address their business imperatives

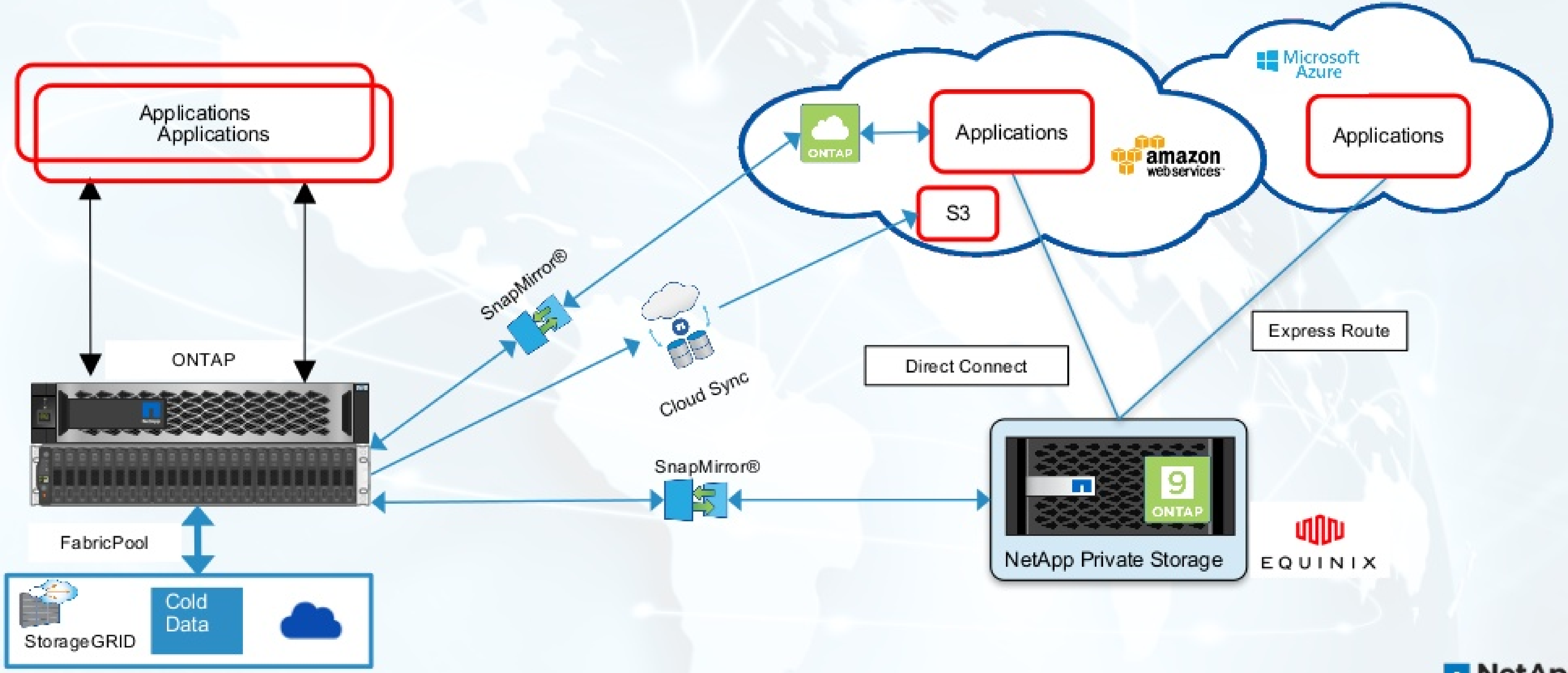


HARNESS the power of the hybrid cloud

BUILD a next-generation data center

MODERNIZE storage through data management

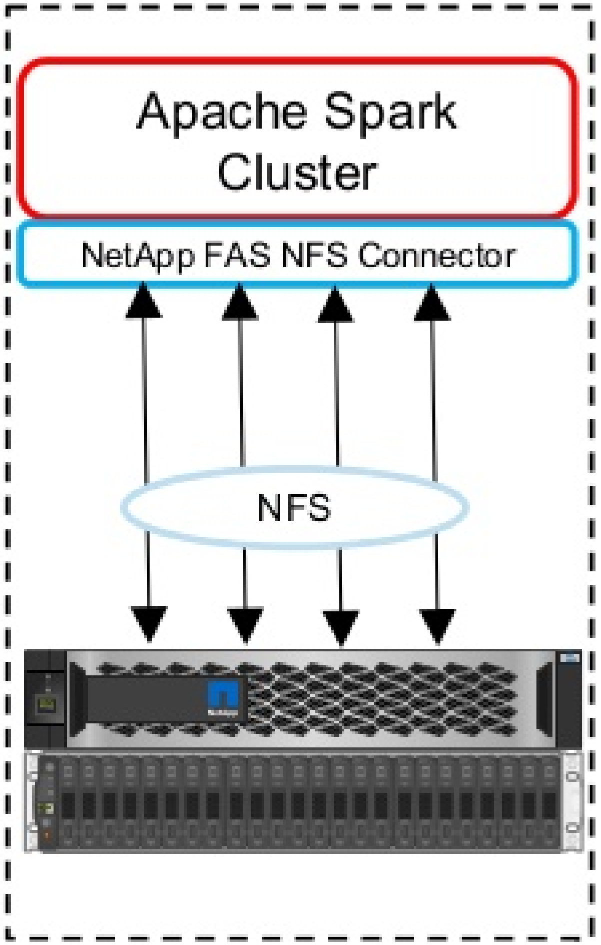
Introducing Data Fabric Building Blocks for Analytics



NetApp

In Place Analytics:

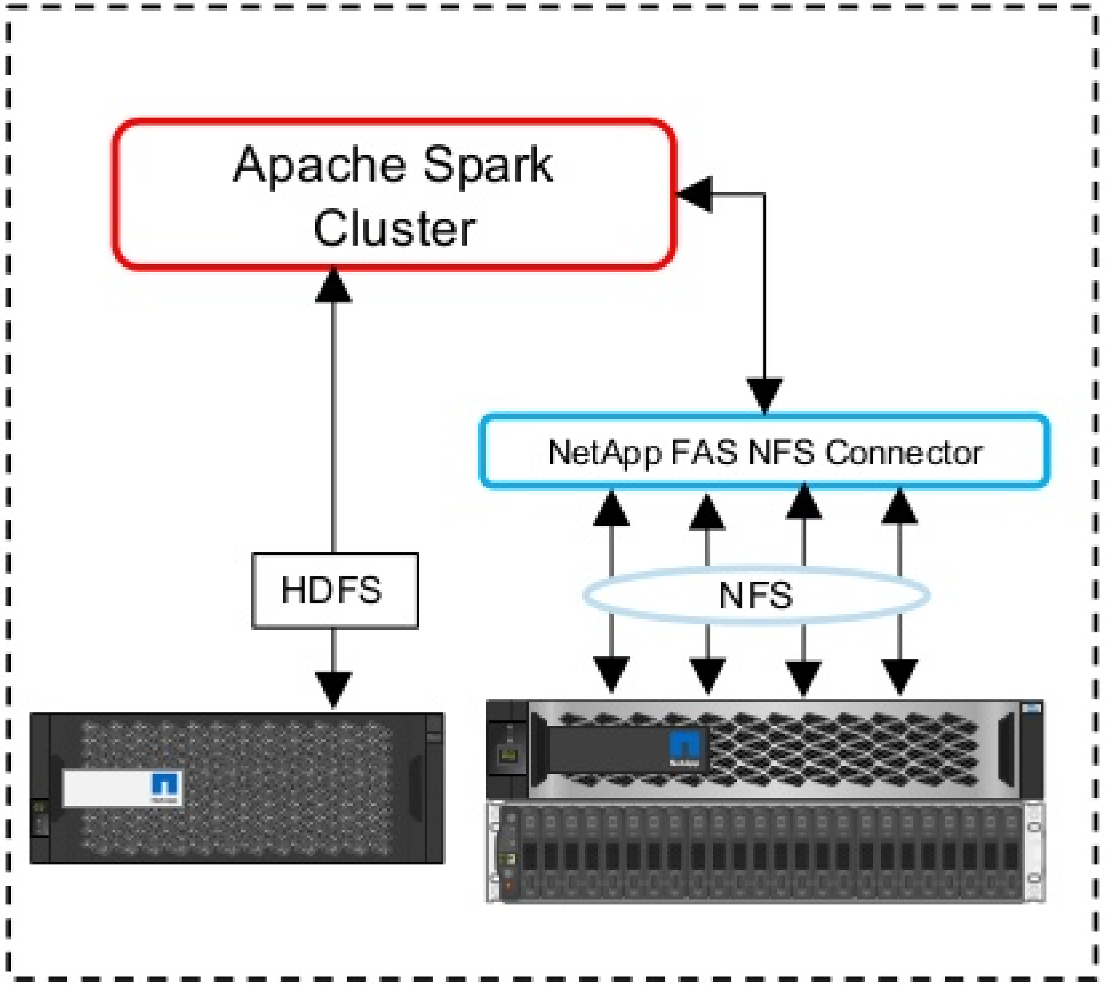
+ Enable big data analytics on NFSv3 data



Confit 1 : NFS as a Storage

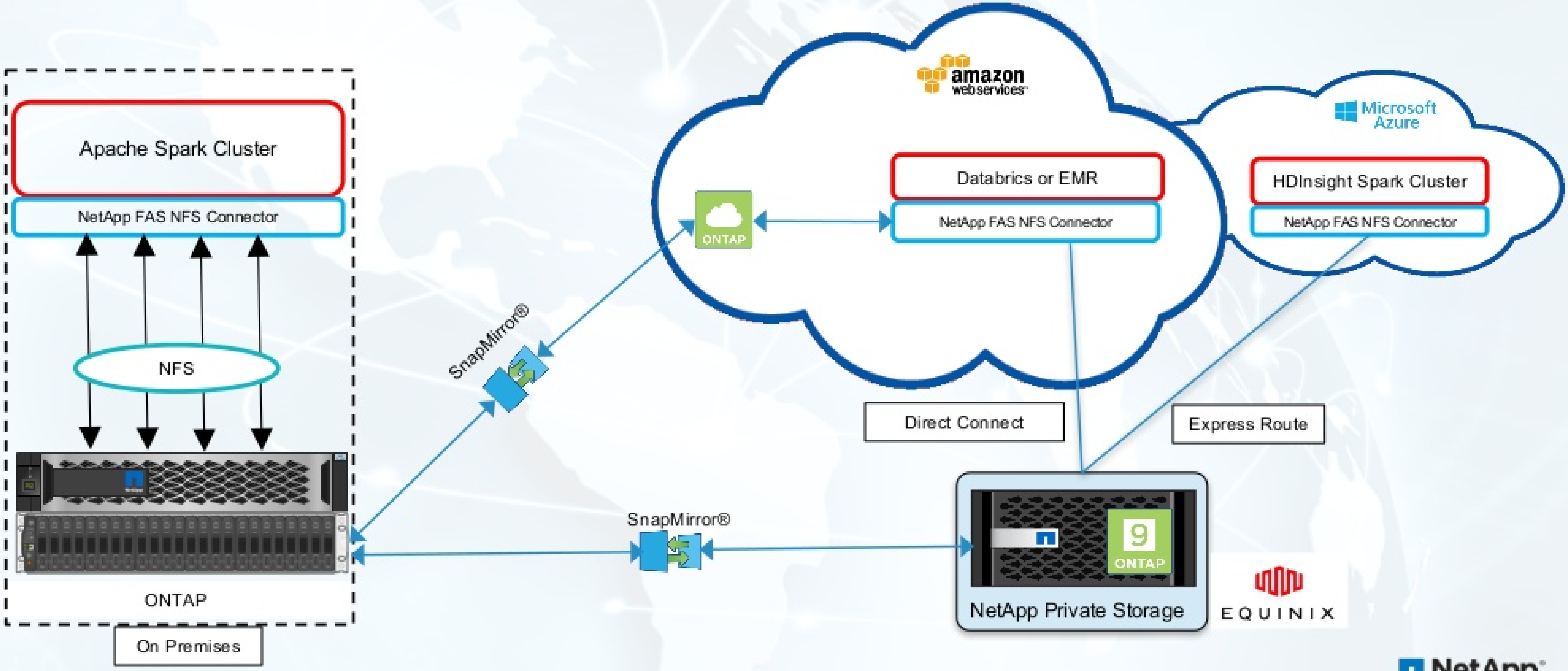
Key Benefits

- Avoid data move to HDFS. Reduce replicas
- Scale compute and storage independently
- Enterprise data protection
- Hybrid cloud deployment
- Hortonworks Certified



Confit 2 : HDFS and NFS in Single Spark Cluster

Analytics with Data Fabric



NetApp

SPARK SUMMIT 2017

IoT Customer Use Cases on Data Fabric

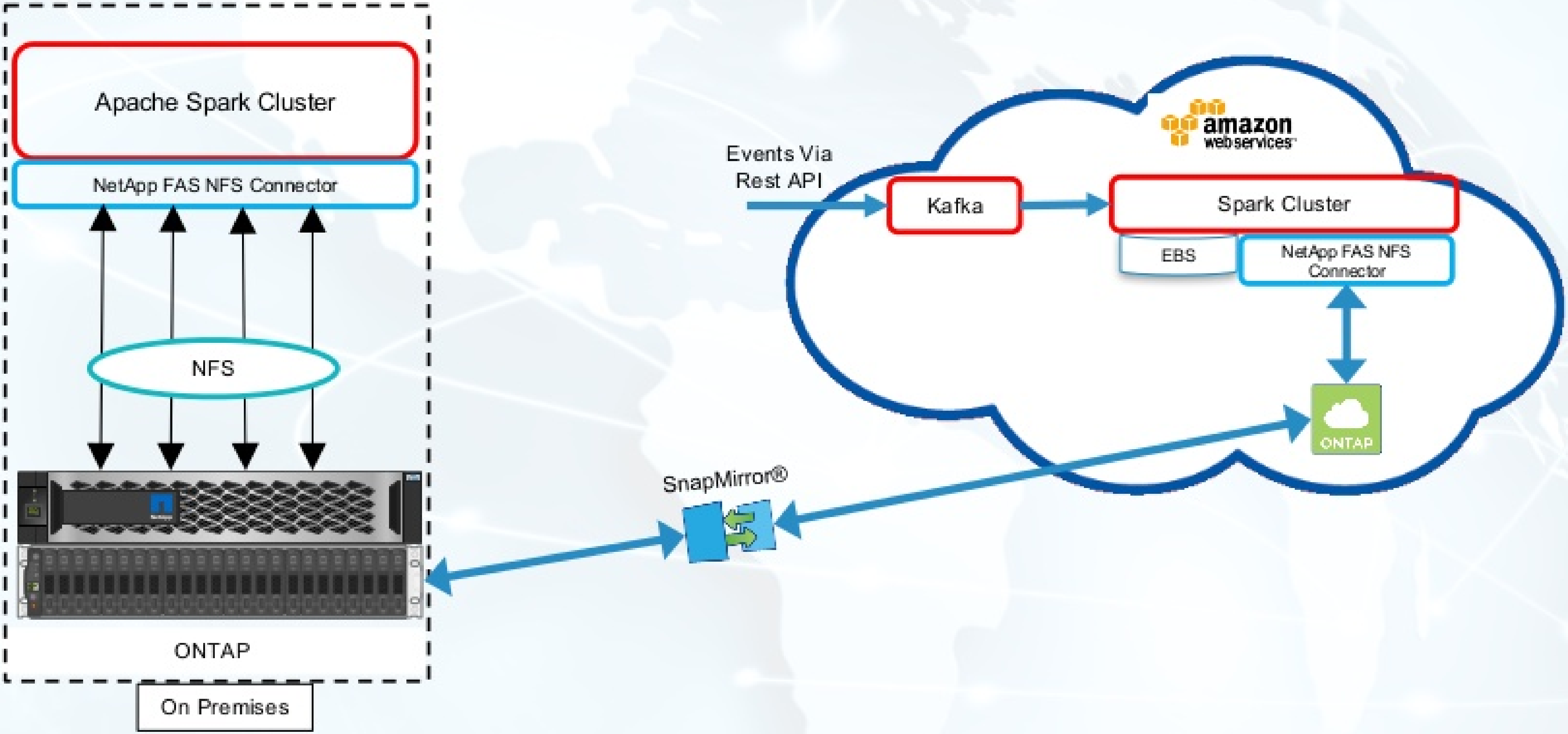
Customer Scenario

Broadcasting Provider

- IoT data received in AWS and analyzed using **Apache Spark cluster** in AWS
- Data Management Challenge:
  - How to Backup 10 TB data without load on cluster?
  - How to protect the data to on-premises?

An Architecture for Processing IoT-Data Ingested in Cloud

Backup and DR to On Premises



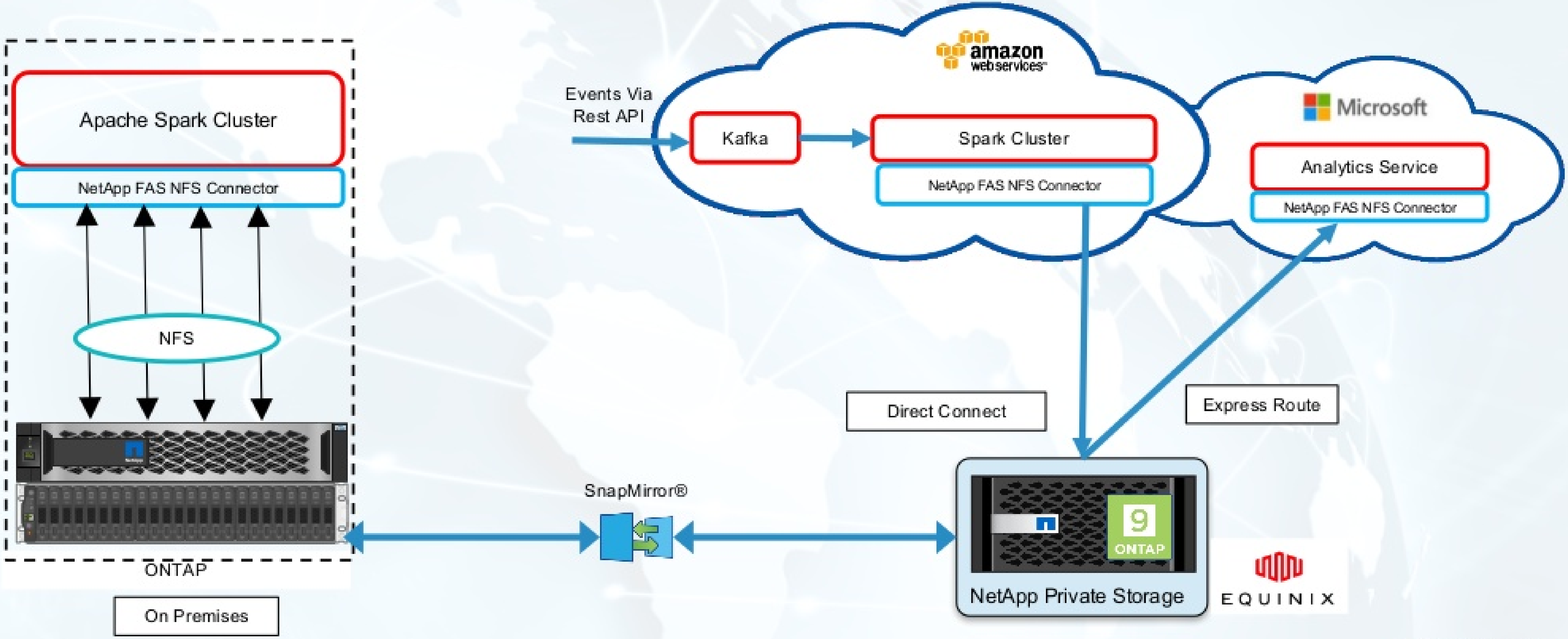
Customer Scenario

IT Service Provider

- IoT data is received in AWS and analyzed using **Apache Spark Cluster** in AWS
- Data Management Challenge:
  - How to reduce the solution cost?
  - How to consume analytics services in data center and multiple clouds?

An Architecture for Processing IoT-Data Ingested in Cloud

Multi Cloud Connectivity



NetApp

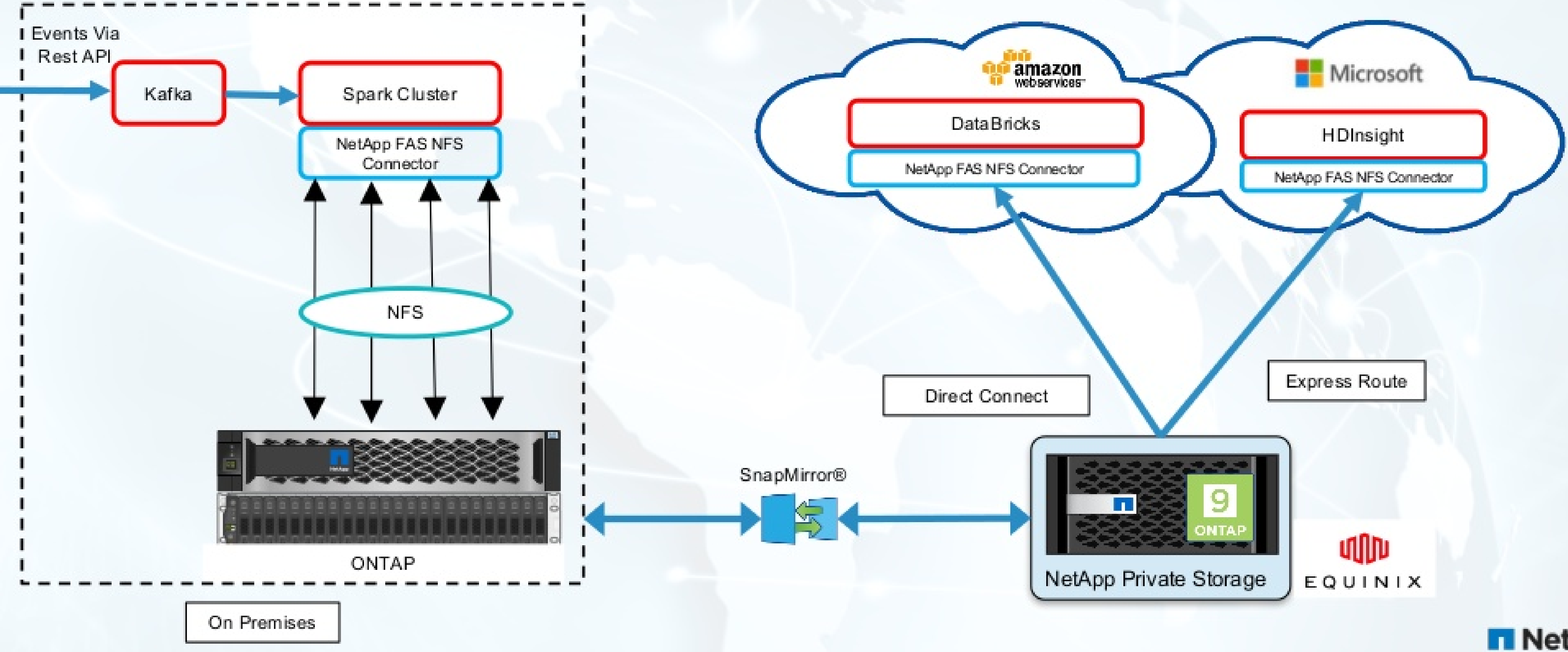
Customer Scenario

Insurance Company

- IoT data received on-premises and analyzed using **Cloudera Spark Cluster** across data center and cloud
- Data Management Challenge:
  - How to leverage cloud computation for analytics ?
  - How to consume legacy data (7PB) for analytics?

An Architecture for Processing IoT-Data Ingested on premises

DR in Cloud; Analytic across data centers



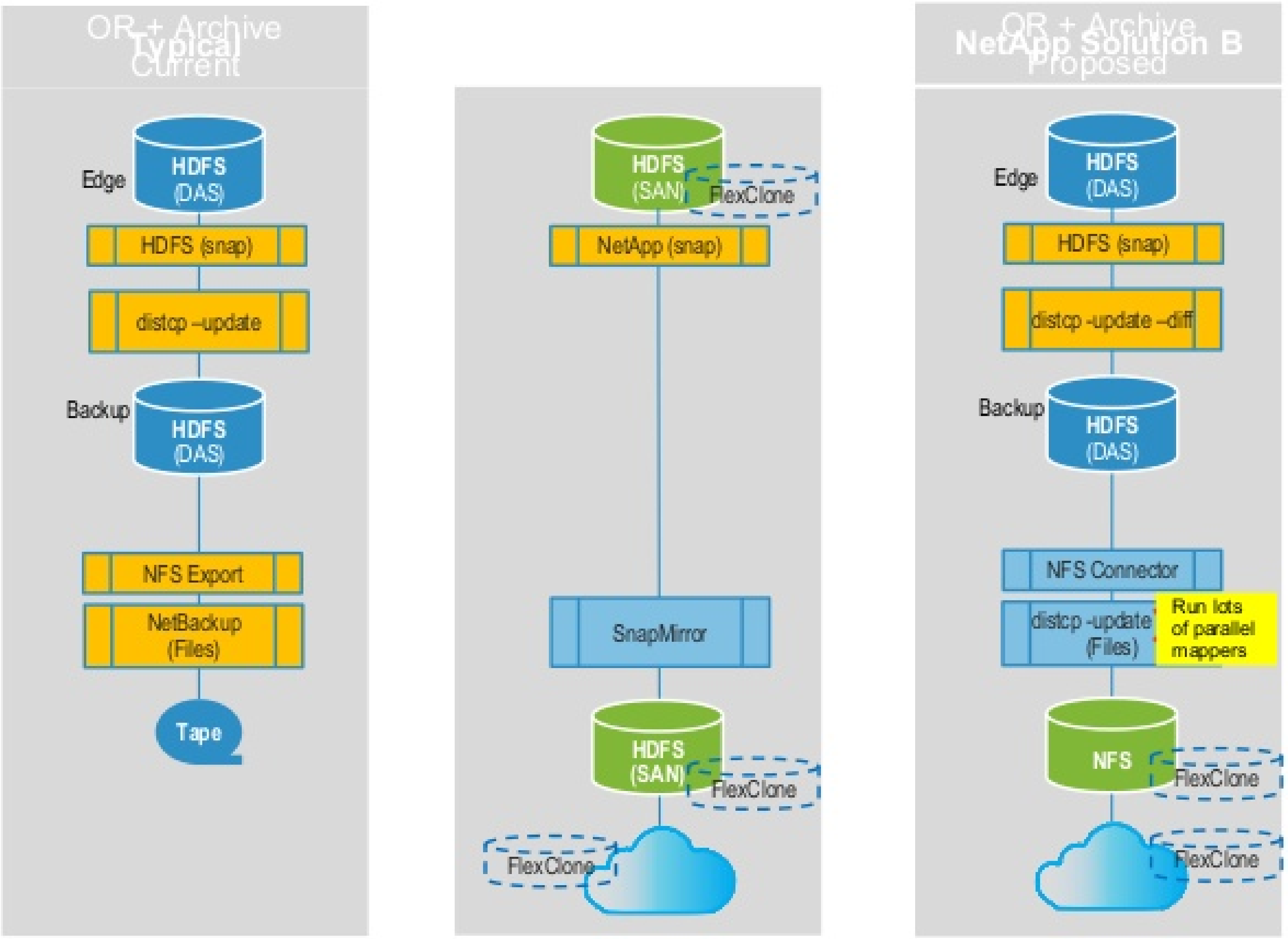
Customer Scenario

Large Bank

- IoT data received on-premises and stored in Hadoop Data Lake. Data needs to be backed up for compliance

- Data Management Challenge:
  - How to reduce the backup window and optimize solution cost?
  - How to minimize impact on analytics performance during backup?

Use Case: Backup for IoT Data



NetApp Backup Solution A

- NetApp Snapshot Backup
- Backup Archival
- Cloud Compatible

NetApp Backup Solution B

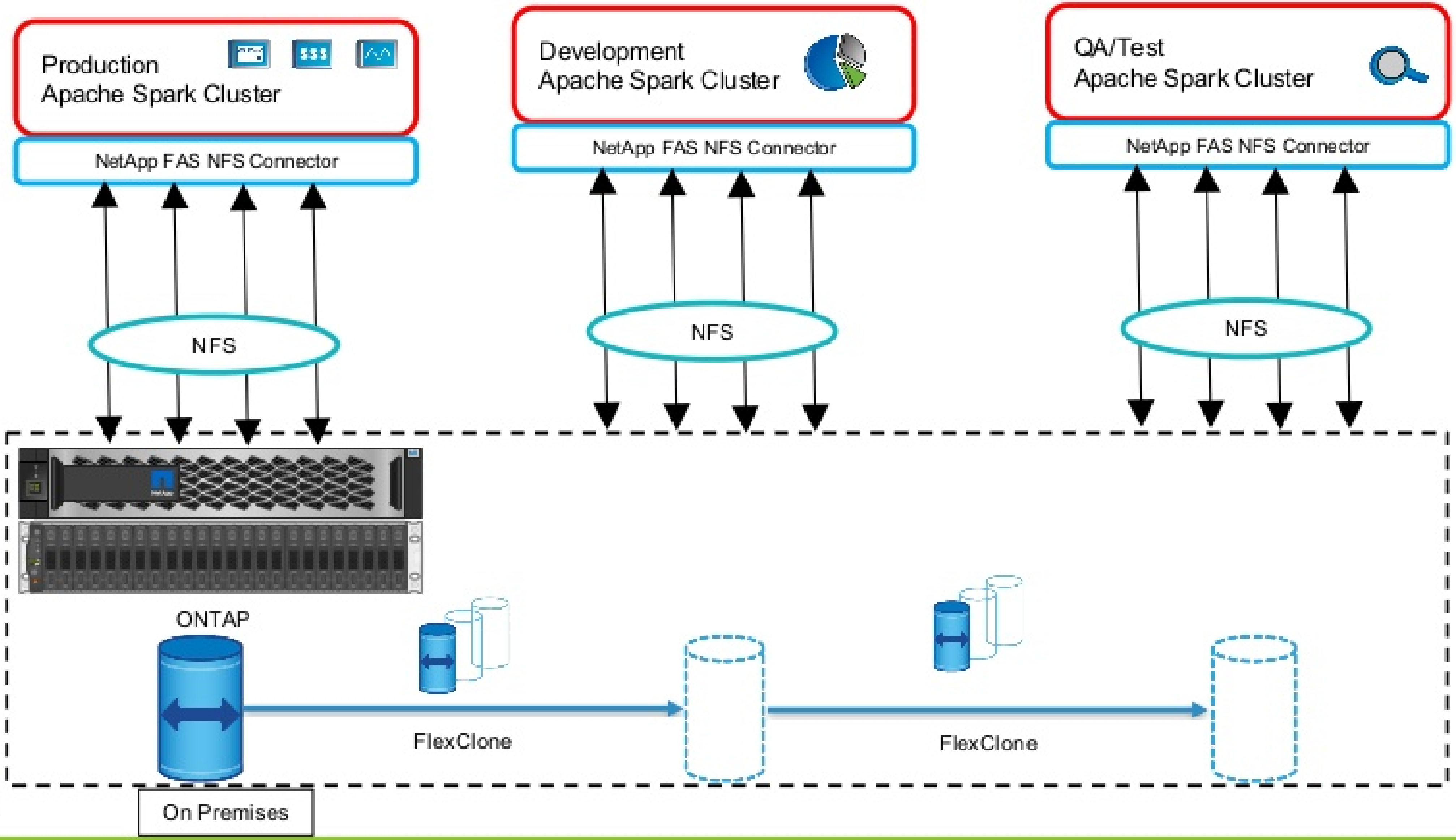
- Hadoop Native Support
- Offload Backup Operation
- Enterprise Management

Customer Scenario

Online Music Distribution

- Large Hadoop Data Lake implementation on premises with Multiple Spark clusters
- Data Management Challenge:
  - How to make data available for dev/test teams?
  - How to build the new cluster in minutes from an existing cluster?

Use Case: Dev/Test for IoT Data



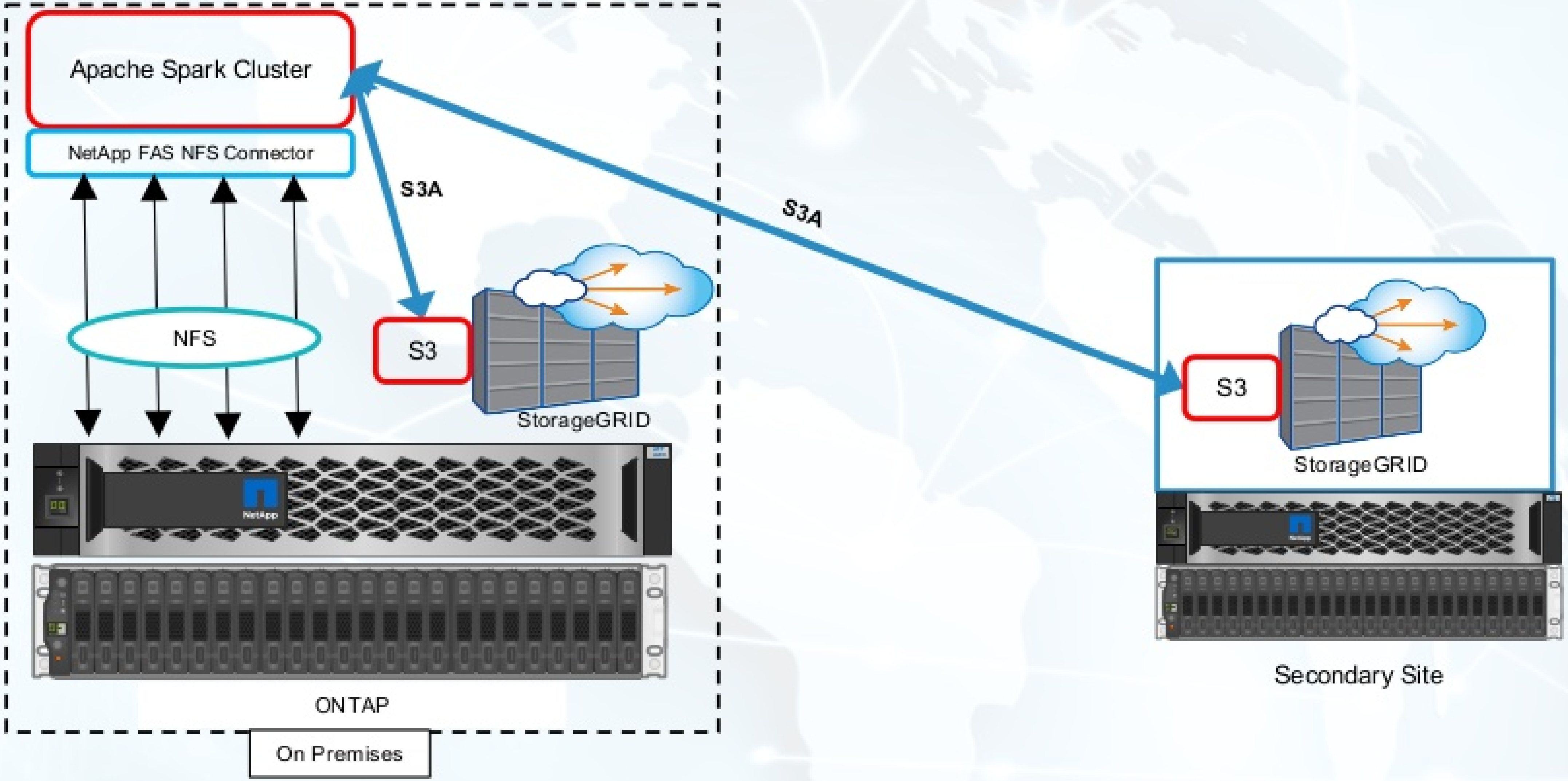
Customer Scenario

Online Marketplace

- Run analytics on archival data in object store
- Data Management Challenge:
  - How to run Hadoop jobs in object store
  - Archive the Hadoop data on primary or a secondary site

Analytics with NetApp StorageGRID

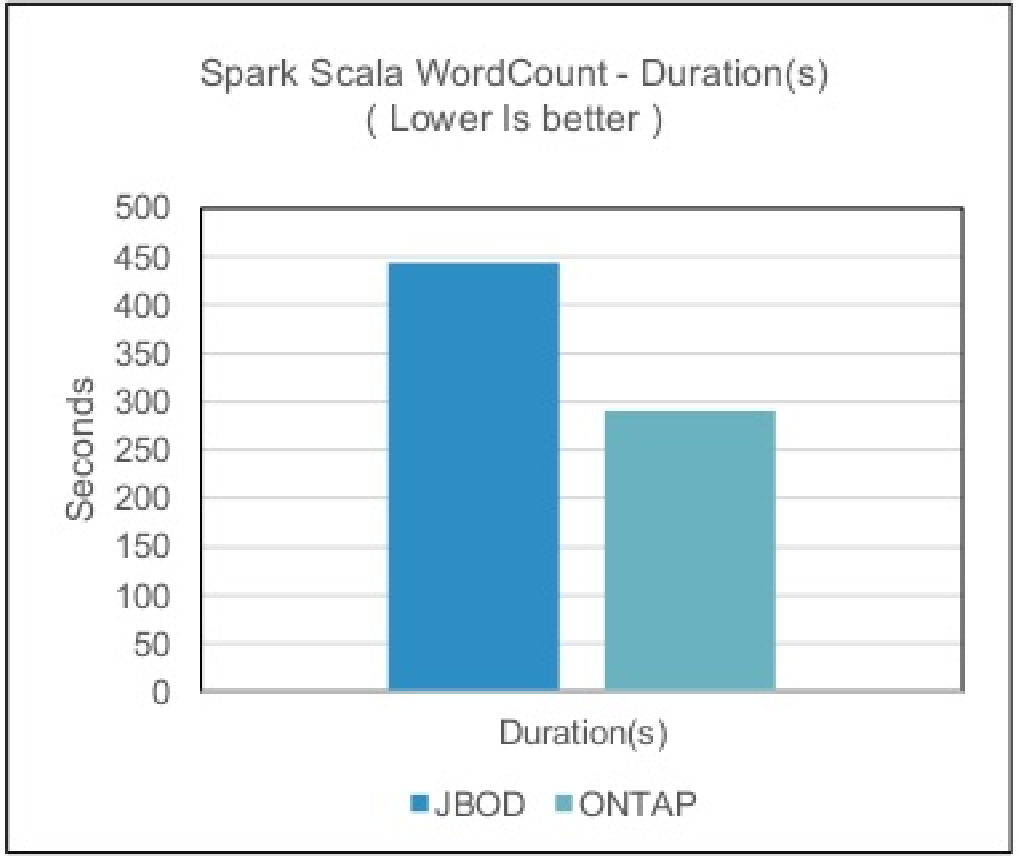
In place analytics with StorageGRID

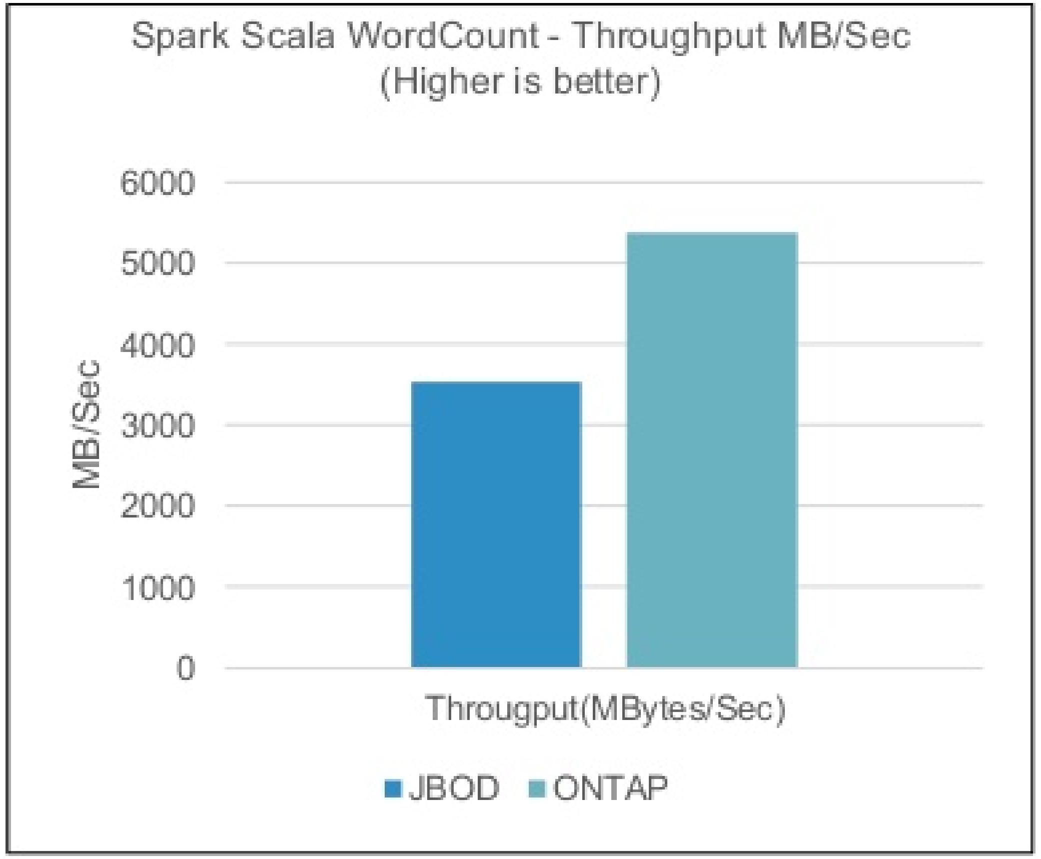


Spark Performance

HiBench – Spark Scala Word Count

Word Count





- Input dataset size – 1.5TB
- ONTAP– 52% better than JBOD

Type	Hadoop Worker Nodes	Drives per Node	Number of Storage Arrays
JBOD	6	12	NA
ONTAP	6	6	1

Key Takeaways

Flexibility and Agility

- On Demand analytics with Hybrid Cloud/Multi Cloud deployments
- Rapid provisioning of clusters for test/dev environments

Lower Cost

- Add storage capacity without adding compute nodes
- One copy vs 3 copies of data for HDFS
- Data Tiering with FabricPool

Enterprise Data Protection

- Efficient backup, DR and Archival

Further Resources

- Please visit us at: Booth #512
- Visit our Big Data Website: www.netapp.com/bigdata

Q & A

SPARK SUMMIT 2017

Thank You.

Nilesh Bagad: nileshb@netapp.com

Karthikeyan Nagalingam: nkarthik@netapp.com