Microsoft

Microsoft Build

May 6–8, 2019

Microsoft

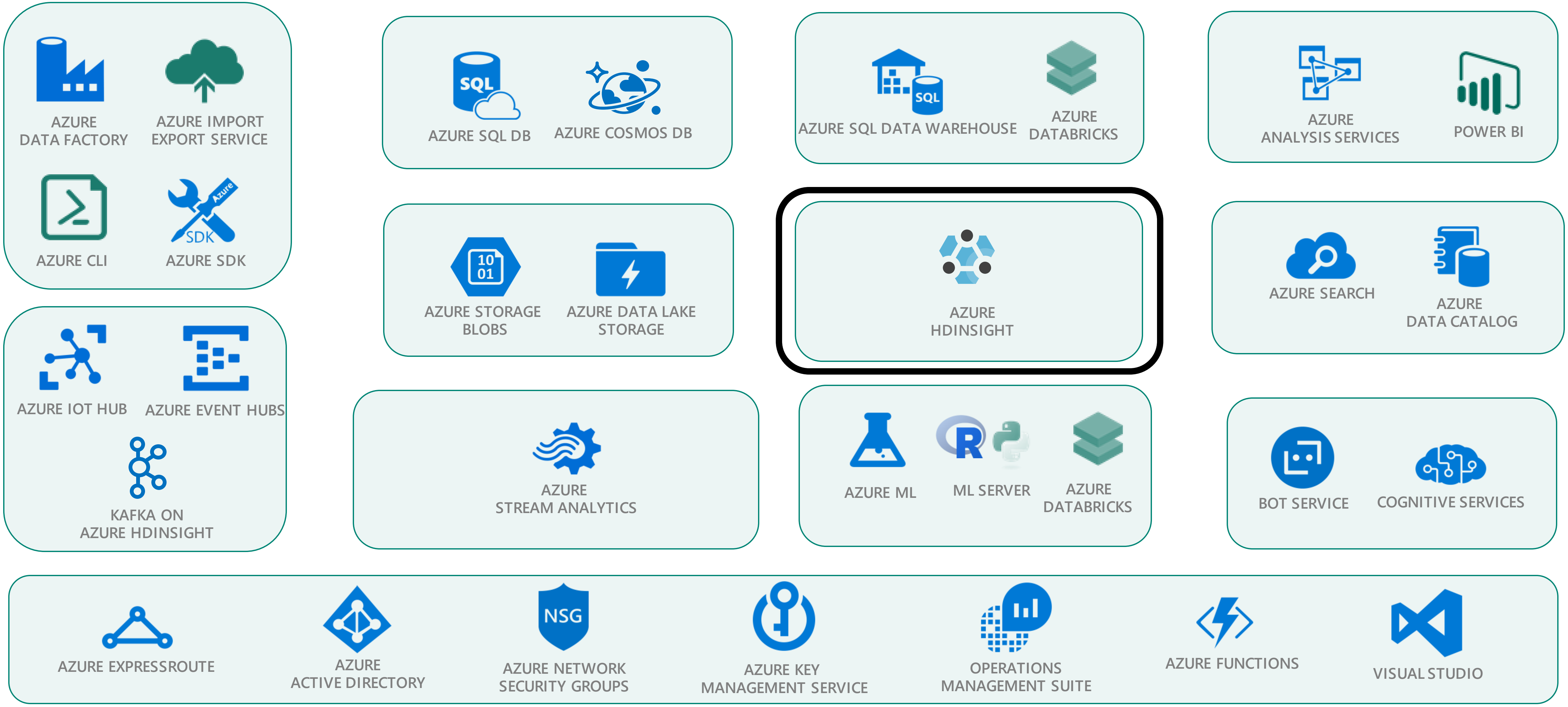
Data Architect's guide for successful Open Source patterns in Azure w/ Spark, Hive, Kafka & HBase.

Ashish Thapliyal
Principal Program Manager, Azure HDInsight
@ashishth

A customer journey!

- Walk through actual customer journey while architecting a large data lake in Azure using HDInsight
- Why HDInsight?
- Not starting from zero, already have Hadoop cluster running on-prem
- Multiple use cases: Batch, Real Time processing, Data science & BI
- Challenges, tradeoffs & tips & tricks

The Azure Data Landscape



AZURE DATA FACTORY

AZURE IMPORT EXPORT SERVICE

AZURE CLI

AZURE SDK

AZURE IOT HUB

AZURE EVENT HUBS

KAFKA ON AZURE HDINSIGHT

AZURE SQL DB

AZURE COSMOS DB

AZURE STORAGE BLOBS

AZURE DATA LAKE STORAGE

AZURE STREAM ANALYTICS

AZURE SQL DATA WAREHOUSE

AZURE DATABRICKS

AZURE HDINSIGHT

AZURE ML

ML SERVER

AZURE DATABRICKS

AZURE ANALYSIS SERVICES

POWER BI

AZURE SEARCH

AZURE DATA CATALOG

BOT SERVICE

COGNITIVE SERVICES

AZURE EXPRESSROUTE

AZURE ACTIVE DIRECTORY

AZURE NETWORK SECURITY GROUPS

AZURE KEY MANAGEMENT SERVICE

OPERATIONS MANAGEMENT SUITE

AZURE FUNCTIONS

VISUAL STUDIO

@ashishth

Azure HDInsight

A secure and managed Apache Hadoop and Spark platform for building data lakes in the Cloud

Open Source

- 100% Apache Open Source
- The most popular open source frameworks
- Part of the Hortonworks HDP distribution

Managed

- 99.9% availability SLA
- Cluster Health Monitoring
- Integration with Azure Log Analytics
- Highly optimized for Azure

Secure & Compliant

- Role based access control
- Azure AD & Kerberos based authentication
- Strong VNET and service endpoint support
- The most trusted and compliant platform

Productive

- Works with the tools developers already have
- Special extensions for advanced debugging and diagnostics

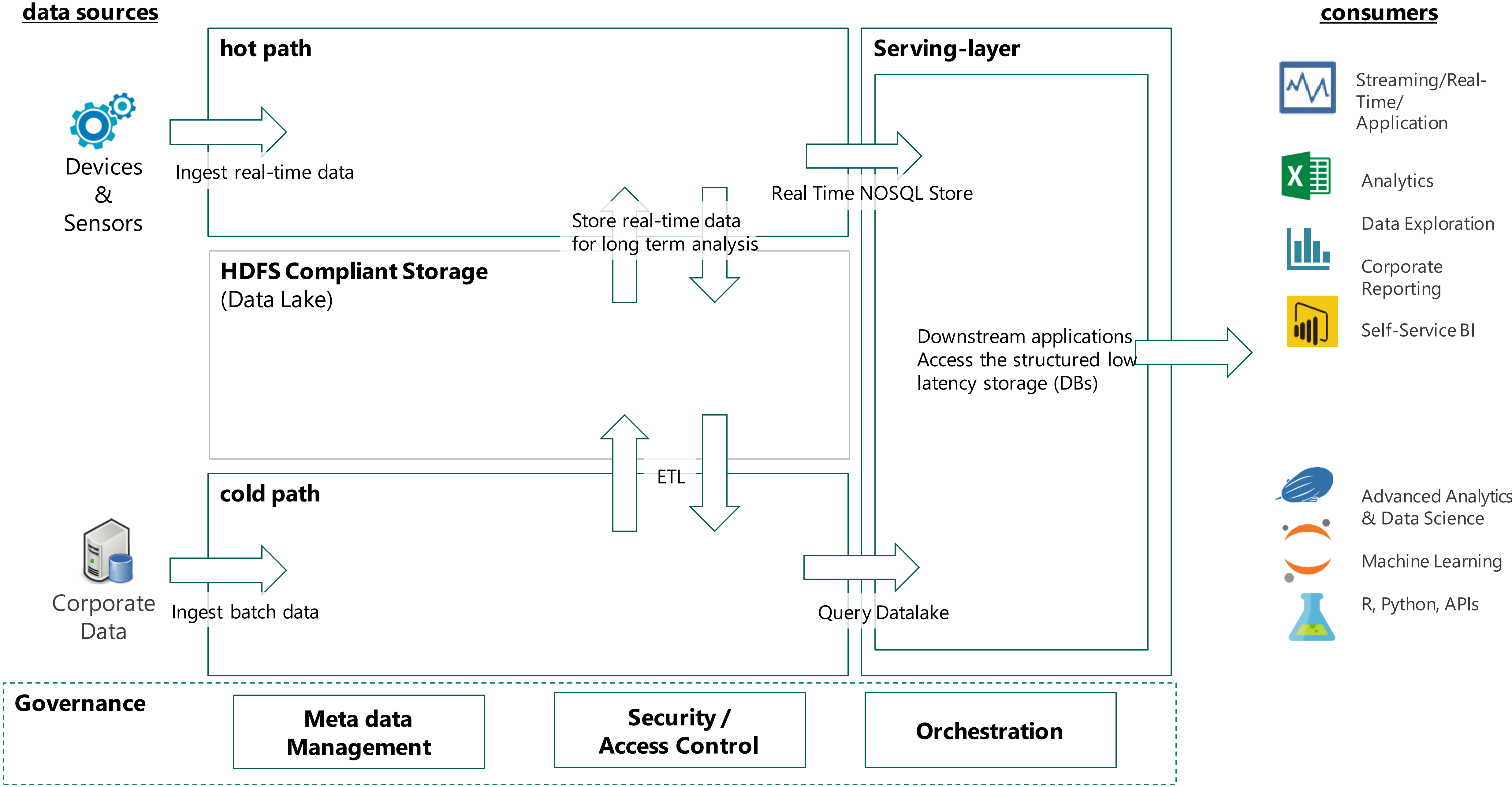
Lift & Shift

- Move workloads from on-prem or other clouds without code changes
- Curated application platform for wide variety of use cases

HIVE kafka APACHE Spark APACHE HBASE APACHE STORM™

@ashishth

Solution architecture



data sources

Devices & Sensors

Corporate Data

hot path

Ingest real-time data

Store real-time data for long term analysis

Real Time NOSQL Store

HDFS Compliant Storage (Data Lake)

ETL

cold path

Ingest batch data

Query Datalake

Serving-layer

Downstream applications Access the structured low latency storage (DBs)

consumers

Streaming/Real-Time/ Application

Analytics

Data Exploration

Corporate Reporting

Self-Service BI

Advanced Analytics & Data Science

Machine Learning

R, Python, APIs

Governance

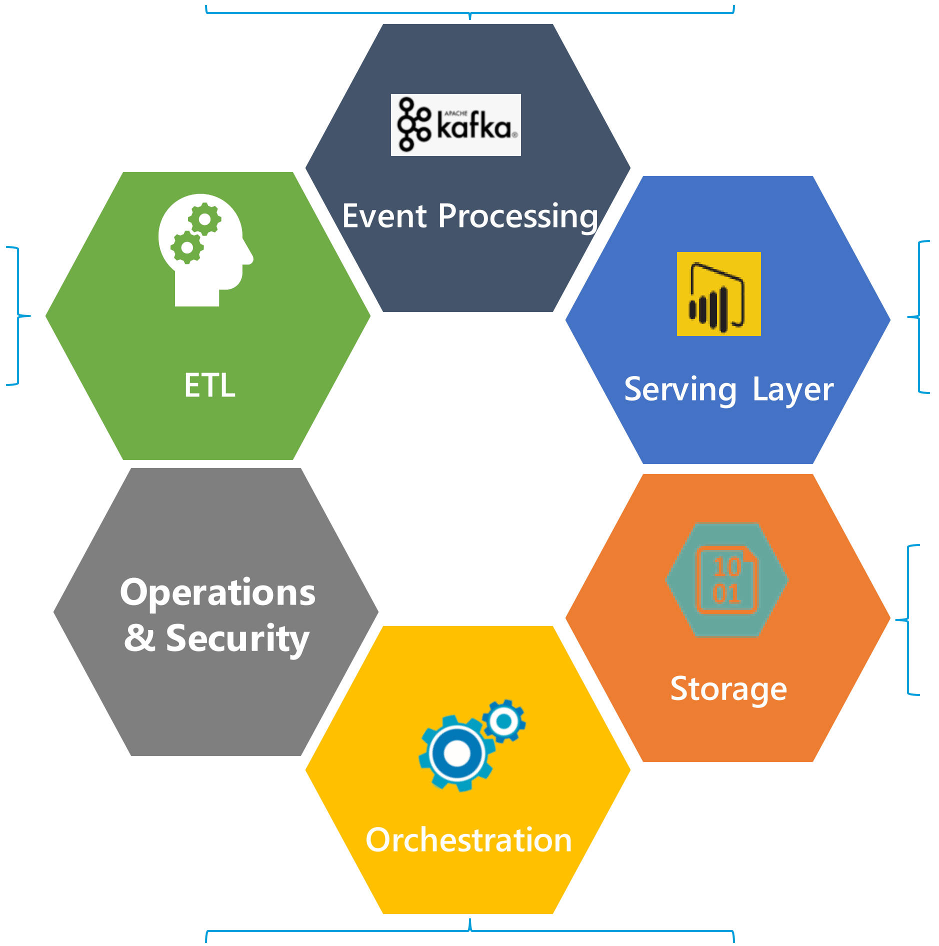
Meta data Management

Security / Access Control

Orchestration

Many things to figure out

Spark Streaming or Storm



Event Processing

Serving Layer

Storage

Orchestration

Operations & Security

ETL

Pig, Hive or Spark

Presto or Hive LLAP?

Which storage system?

ADF/Airflow or Oozie?

OSS Framework choices & tradeoffs

1. ETL technology choices

	Spark	Pig	Hive
Designed for	ETL	ETL	Data warehousing
Adoption	High, increasing	Low, decreasing	Stable
Number of connectors	Highest	High	High
Languages	Python, R, Scala, Java, SQL	Pig	SQL
Performance	High	Medium	Medium

2.Streaming engine technology choices

	Spark Structured Streaming	Storm
Adoption	High, increasing	Decreasing
Event processing guarantee	Exactly once	At least once
Throughput	High	Low
Processing Model	Micro Batch	Real-Time
Latency	High	Low
Event time support	Yes	Yes
Languages	Python, R, Scala, Java, SQL	Java

3.Interactive Query technology choices

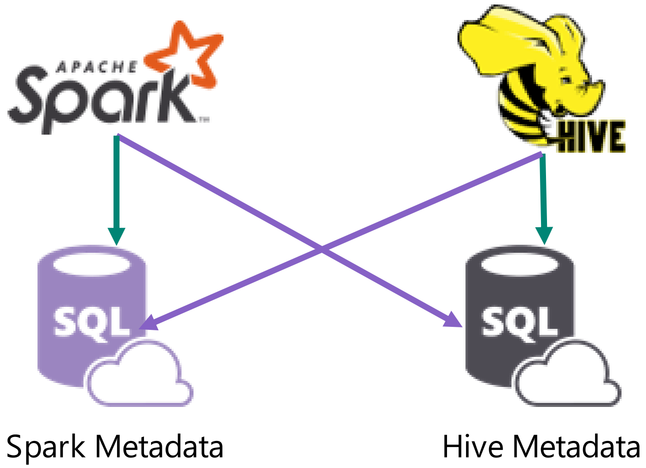
Capability	Hive LLAP	Spark SQL	Presto
Interactive Query Speed	High	High	Medium
Scale	High	High	Low
Caching	Yes	Yes	Early Support
Result Caching	Yes	No	No
Intelligent Cache Eviction	Yes	No	No
Materialized Views	Yes	No	No
Complex Fact to Fact Joins	Yes	Yes	No
Transactions	Yes	No	No
Query Concurrency	High	Low	Low
Row , Column level security	Yes [Apache Ranger+ AAD]	Medium	Medium
Rich end user Tools	Yes	Yes	Yes
Language Support	SQL, UDF	SQL, Scala, Python	SQL
Data Source Connector	Storage Handlers	Data Sources	High number of

@ashishth

How about Metastore?

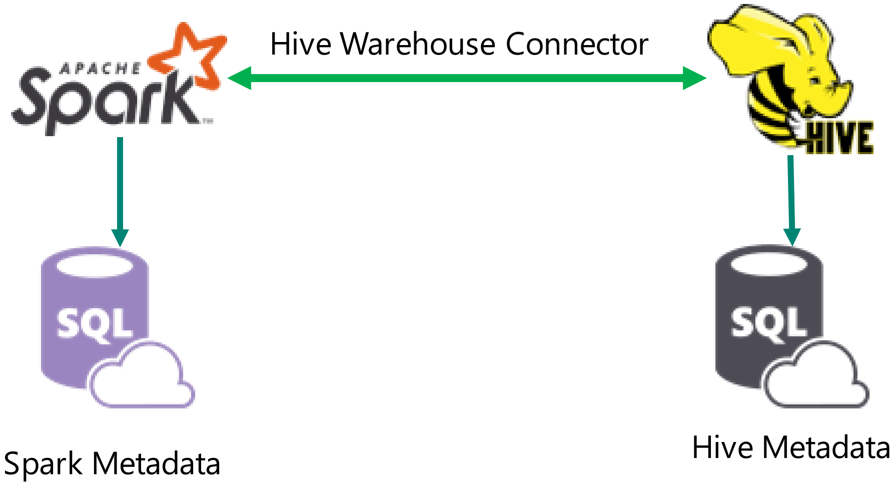
Tip: Spark & Hive Metastore

Azure HDInsight 3.6 with Hadoop 2.6



- Spark executors talk directly to Hive Metastore
- Reliability and compatibility issues
- Cannot take advantage of the native query engine

Azure HDInsight 4.0 with Hadoop 3.x



- **New Hive Warehouse Connector**
- Apache Arrow based communication between Spark executors and Hive LLAP
- Smart predicate pushdown
- Transactional access to Hive tables from Spark

Hive Metastore migration tool: https://azure.microsoft.com/en-us/blog/hdinsight-metastore-migration-tool-open-source-release-now-available/

@ashishth

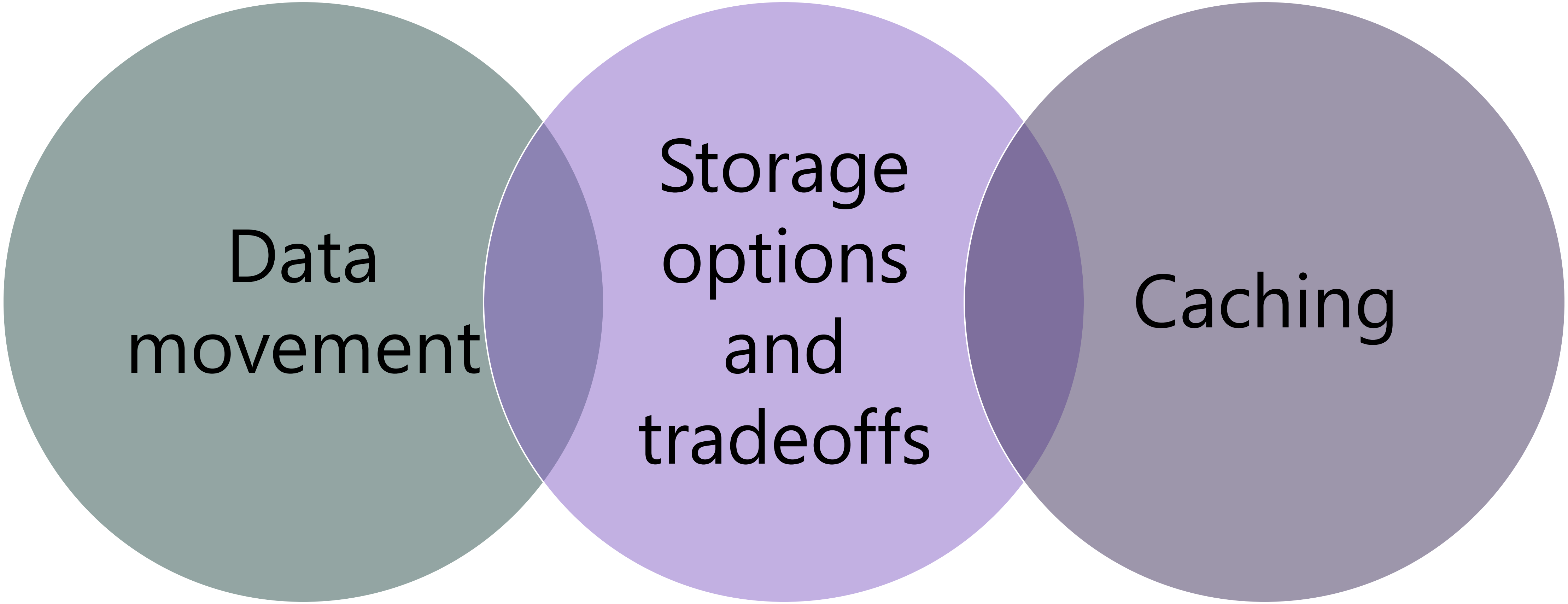
4.Data pipeline orchestration technology choices

	ADF	Airflow	Oozie
Service management	Azure PaaS	IaaS VM	HDInsight
Code	JSON	Python	Java
GUI	ADF V2 has great UX	Good UX	Below Average UX
Community	Microsoft	Growing (10893 Stars)	Declining (454 Stars)
On-demand clusters	Yes	No, but extensible	No
Extensibility	Custom action-only	Full, graph + actions	Custom action-only
Pipeline definition	JSON/UX	Python/ UX	XML/UX
Devops-first design	Yes	Yes	Yes
Pipeline monitoring	Yes	Yes	Yes
Scheduling	Event, Time	Event	Event, Time

@ashishth

Storage & Security

Storage selection: 3 key topics



Data movement

Storage options and tradeoffs

Caching

Data Qty	Network Bandwidth		
	45 Mbps (T3)	100 Mbps	1 Gbps
1 TB	2 days	1 day	2 hours
10 TB	22 days	10 days	1 day
35 TB	76 days	34 days	3 days
80 TB	173 days	78 days	8 days
100 TB	216 days	97 days	10 days
200 TB	1 year	194 days	19 days
500 TB	3 years	1 year	49 days
1 PB	**6 years**	**3 years**	**97 days**
2 PB	12 years	5 years	194 days

Storage Transfer options

Network Transfer with TLS

- Over Internet
- Express Route
- Data Box online Transfer

Shipping data offline

- Data Box offline data transfer

Azure Data Box: offline transfer options

Available for small, medium, or large migrations

PRODUCTS	CAPACITY	DESCRIPTION
Data Box Disk	8 TB, up to 40 TB	USB 3.1 SSD disks Order up to 5 in each pack
Data Box	100 TB	Ruggedized, self-contained appliances
Data Box Heavy	1 PB	

Storage Options with HDInsight

	Type	Latency (Consistency of latency)	Workloads	Bandwidth	Key Benefits
ADLS Gen 2	Hierarchical	10-50ms (Medium)	HDInsight 3.6 & 4.0	Unconstrained	Atomic Rename, File Folder level ACL's
Standard BLOB	Object Store	10-50ms (Medium)	HDInsight 3.6 & 4.0	Unconstrained	Mature
Premium BLOB	Object Store	~5ms (High)	HBase in Preview	Unconstrained	Fast
Premium Managed Disks	Hierarchical	~5ms (High)	Kafka, HBase in preview	Based on disk	Consistent latency
ADLS Gen 1	Hierarchical	10-100ms (Low)	HDInsight 3.6(No HBase)	High	Atomic Rename, File Folder level ACL's

Storage Options with HDInsight

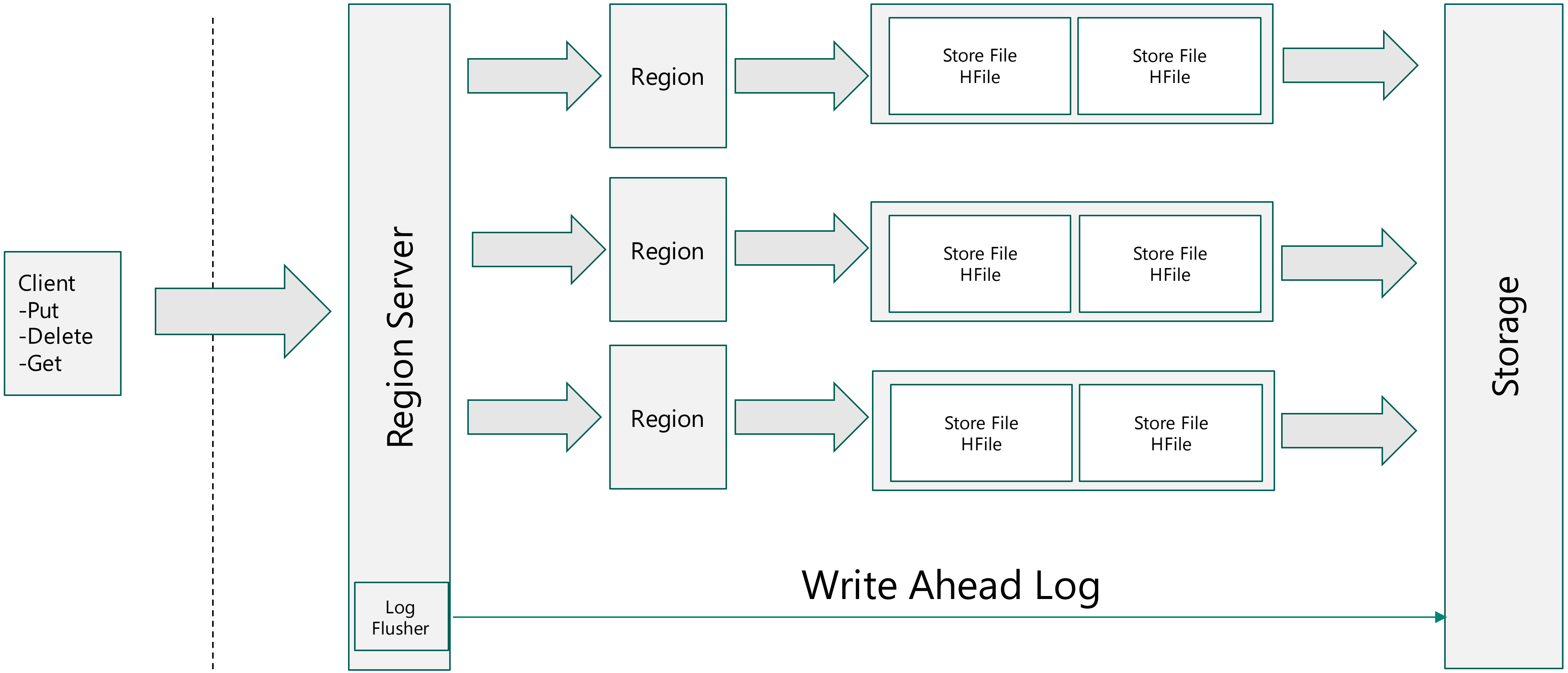
	Type	Latency (Consistency of latency)	Workloads	Bandwidth	Key Benefits
ADLS Gen 2	Hierarchical	10-50ms (Medium)	HDInsight 3.6 & 4.0	Unconstrained	Atomic Rename, File Folder level ACL's
Standard BLOB	Object Store	10-50ms (Medium)	HDInsight 3.6 & 4.0	Unconstrained	Mature
Premium BLOB	Object Store	~5ms (High)	HBase in Preview	Unconstrained	Fast
Premium Managed Disks	Hierarchical	~5ms (High)	Kafka, HBase in preview	Based on disk	Consistent
~~ADLS Gen 1~~	~~Hierarchical~~	~~10-100ms (Low)~~	~~HDInsight 3.6(No HBase)~~	~~High~~	~~Atomic Rename, File Folder level ACL's~~

Don't use ADLS Gen 1 for any new projects

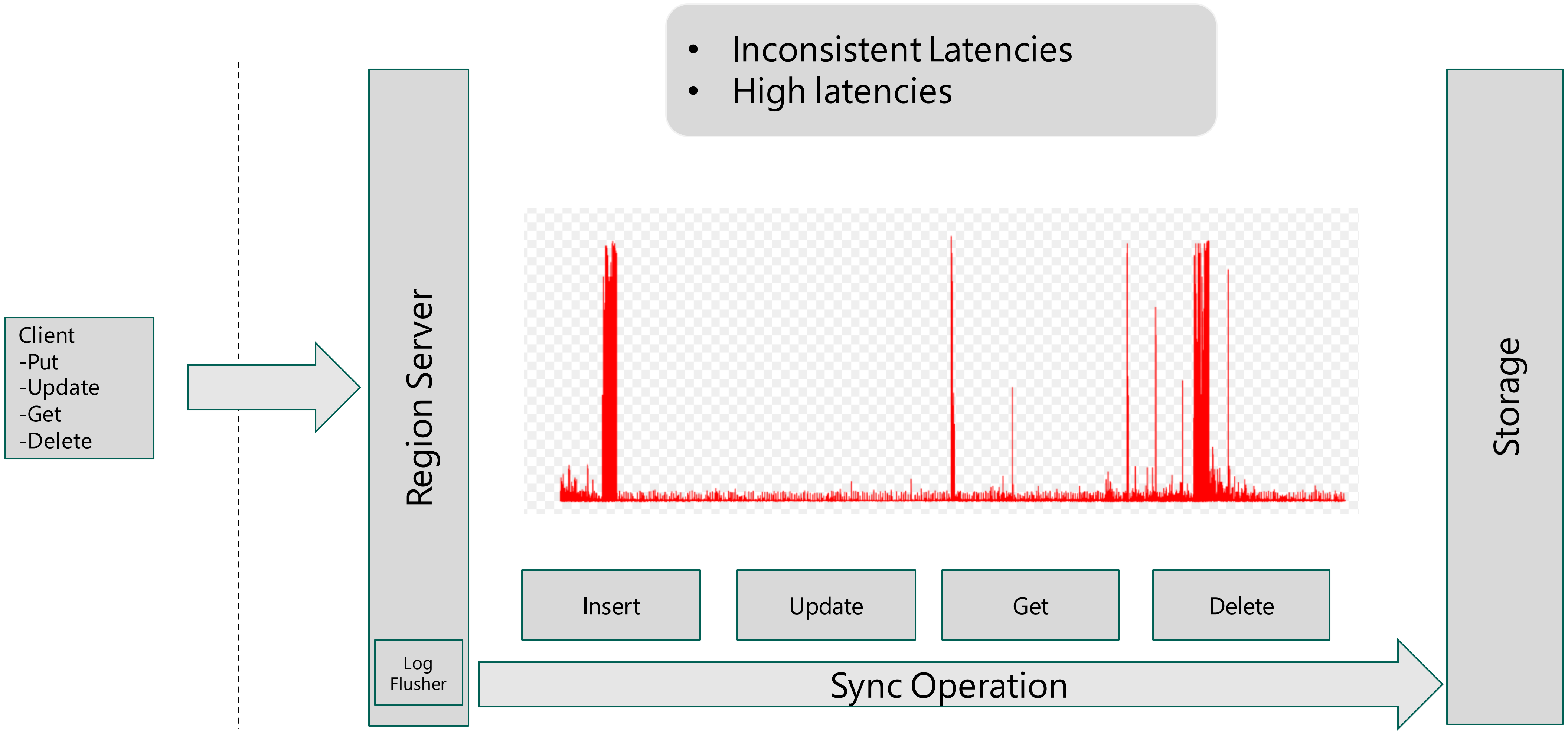
@ashishth

Low latency small writes (HBase use case)

Low latency workload HBase/ Small write



@ashishth



- Inconsistent Latencies
- High latencies

Client
-Put
-Update
-Get
-Delete

Region Server

Log Flusher

Insert

Update

Get

Delete

Sync Operation

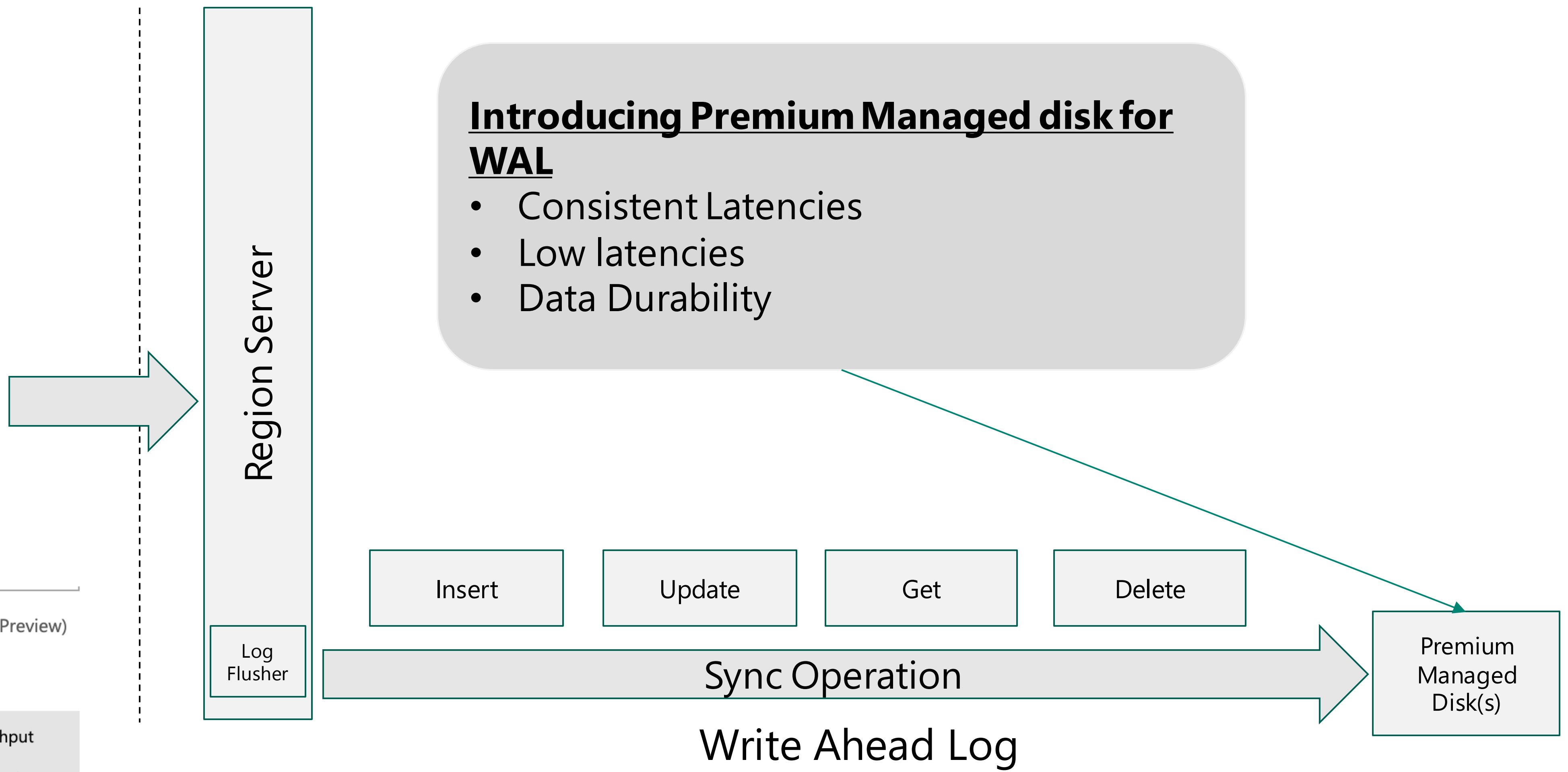
Storage

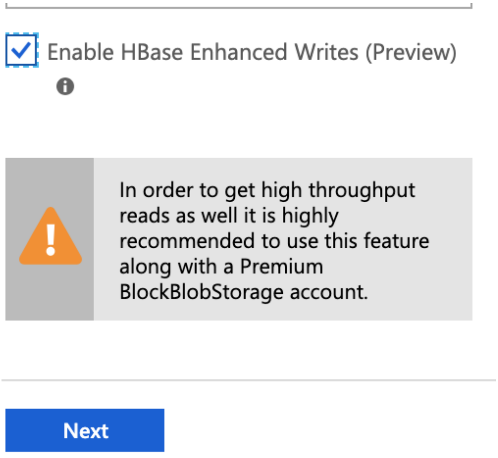
Remote store write path challenges with Write Ahead Log

Introducing Premium Managed disk for WAL

- Consistent Latencies
- Low latencies
- Data Durability

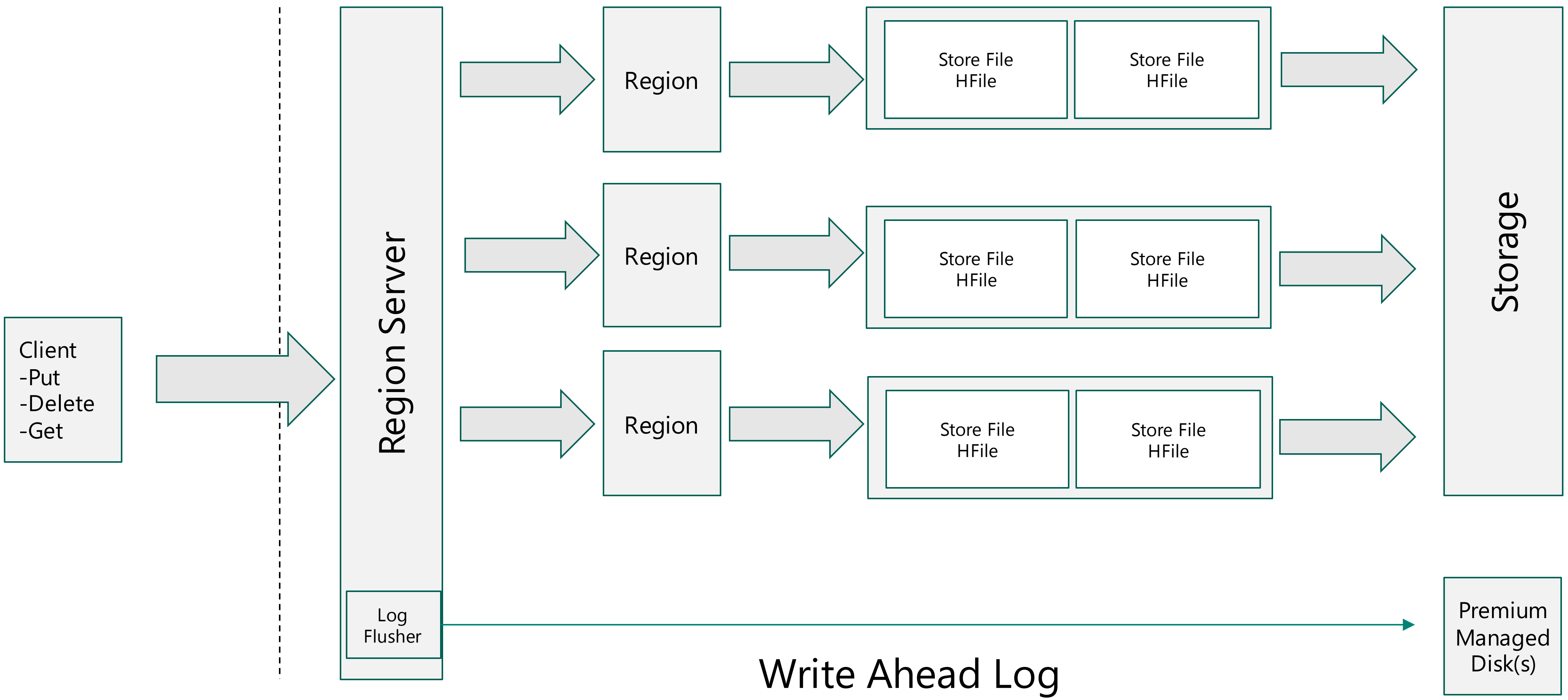
Client
-Put
-Update
-Get
-Delete




☑ Enable HBase Enhanced Writes (Preview)

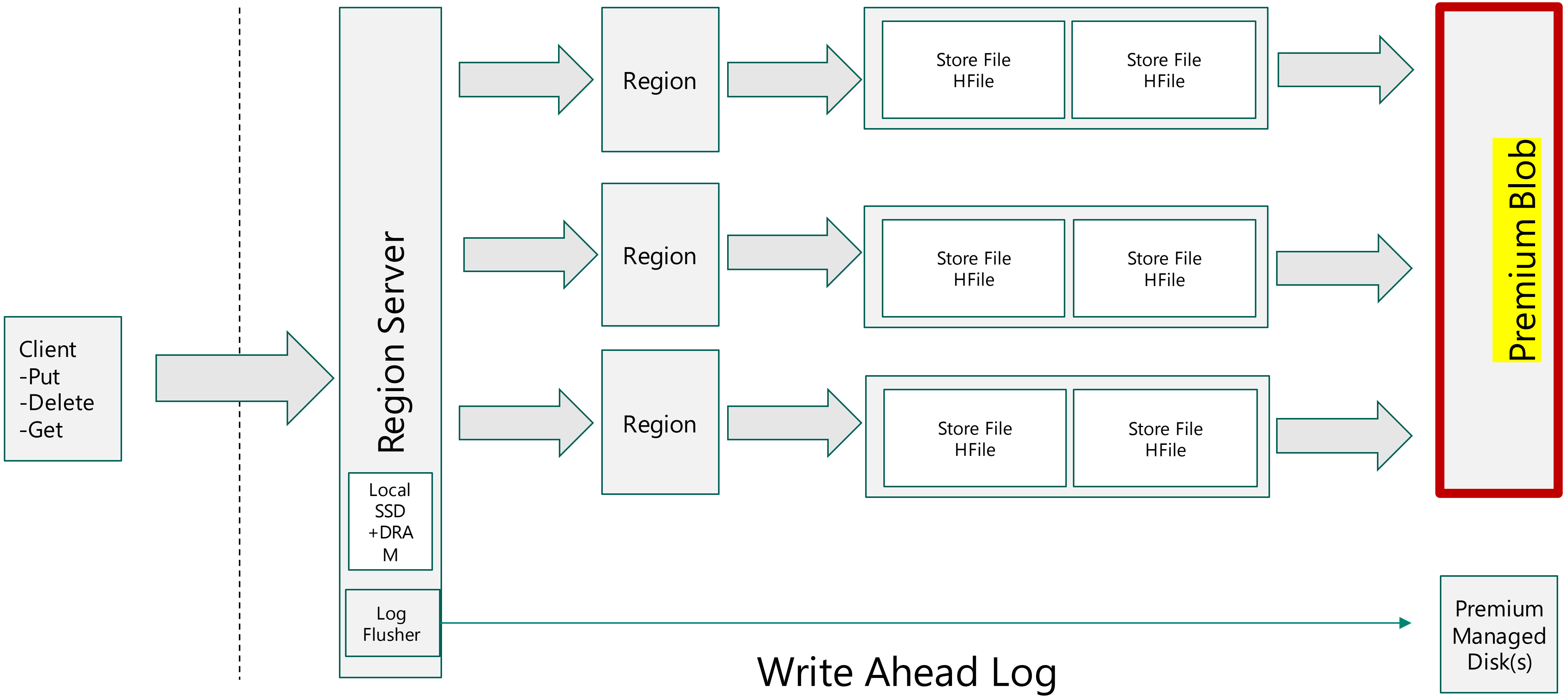
In order to get high throughput reads as well it is highly recommended to use this feature along with a Premium BlockBlobStorage account.

@ashishth


Client
-Put
-Delete
-Get
Region Server
Log Flusher
Region
Region
Region
Store File HFile
Store File HFile
Store File HFile
Store File HFile
Store File HFile
Store File HFile
Storage
Premium Managed Disk(s)
Write Ahead Log

@ashishth

How about Reads?



Introducing support for Premium Blob

@ashishth

Performance (YCSB)

Cluster Type	Operation	Row Size	# ops	#Region Servers	Region Server Node Size	#Clients	Throughput	Avg Latency (ms)	Run Time (min)
Standard	Write	1KB	107,374,182	4	Standard_D4_V2	2	37,958	**0.417**	47
Premium WAL	Write	1KB	107,374,182	**4**	Standard_D4_V2	2	57,812	0.271	31
Standard	Small Write	100 Bytes	1,073,741,824	**4**	Standard_DS4_V2	2	**84,910**	0.186	210
Premium WAL	Small Write	100 Bytes	1,073,741,824	**4**	Standard_DS4_V2	2	**701,234**	0.016	25
Standard	Read	100 Bytes	925,075	4	Standard_D4_V2	2	256	**62**	60
Premium WAL & Premium Blob	Read	100 Bytes	33,503,676	4	Standard_D4_V2	2	9,306	1.7	60
Standard	Large Read	1K	945,682	4	Standard_D4_V2		262	**61**	60
Premium WAL & Premium Blob	Large Read	1K	24,846,209	4	Standard_D4_V2	2	6901	**2.3**	60

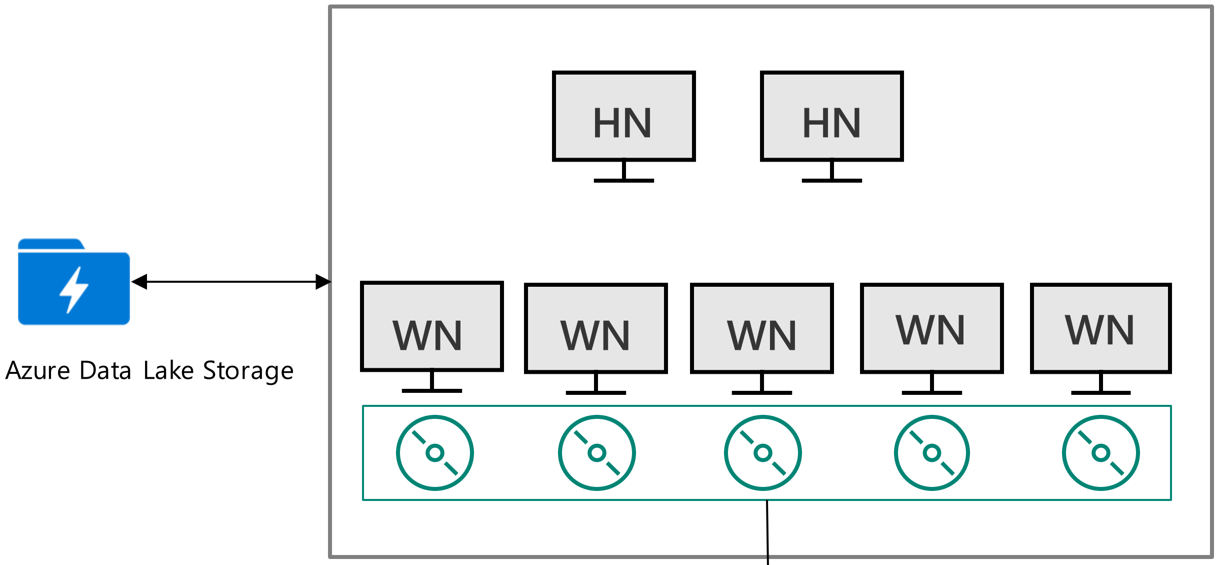
@ashishth

Remote Storage: Caching considerations

Remote Storage: Caching Options

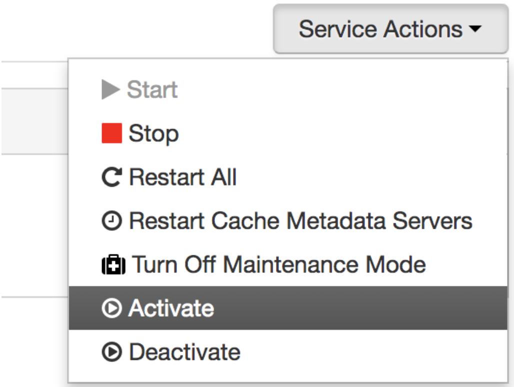
Workload	Caching Options	Key benefits
Spark	Spark IO Cache	Up to ~8 to 10x perf improvements
HBase & Phoenix	Bucket cache	Up 5-10x perf gains on recently read or written data
Hive + LLAP	LLAP Intelligent cache/Result Cache	Up to ~4-100X gain on cached data

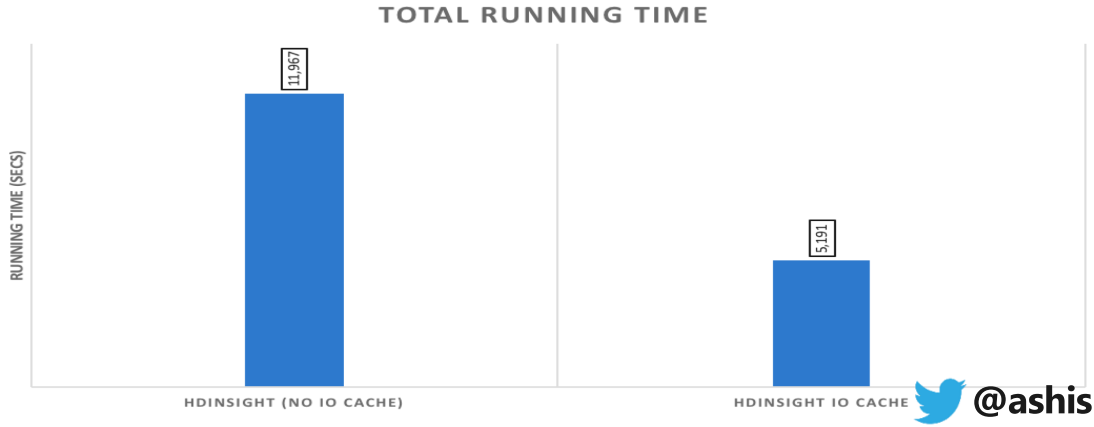
HDInsight IO Cache



INSTANCE	CORE	RAM	TEMP SSD
D1 v2	1	3.50 GiB	**50 GiB**
D2 v2	2	7.00 GiB	**100 GiB**
D3 v2	4	14.00 GiB	**200 GiB**
D4 v2	8	28.00 GiB	**400 GiB**
D5 v2	16	56.00 GiB	**800 GiB**

- Significant Spark performance speed up with IO cache (up to 9X perf gains)
- Automatic cache resource management
- DRAM + Temp SSD makes large cache

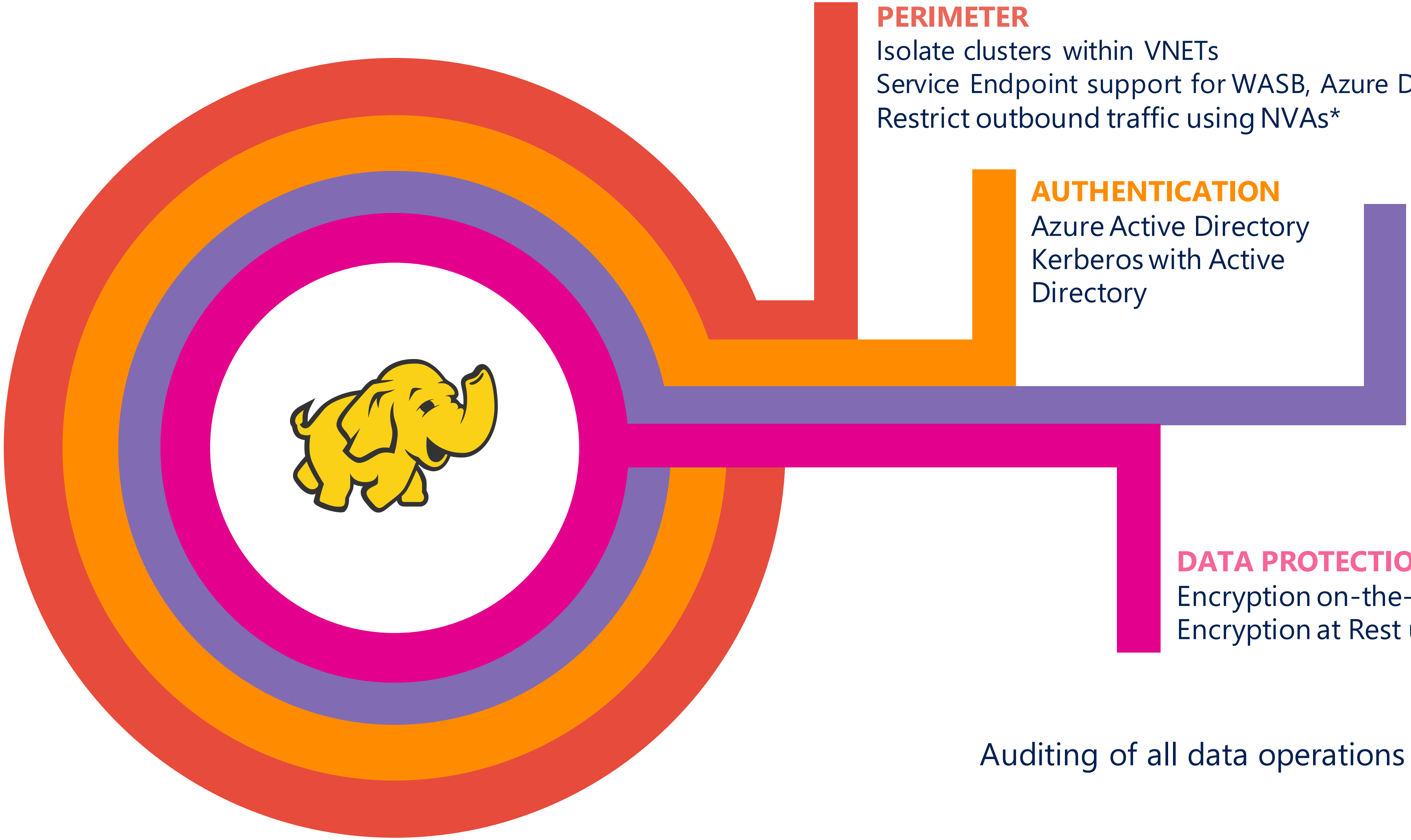




Security

Azure HDInsight: Enterprise Grade Security

Defense in Depth



PERIMETER

Isolate clusters within VNETs
Service Endpoint support for WASB, Azure DB, Cosmos DB
Restrict outbound traffic using NVAs*

AUTHENTICATION

Azure Active Directory
Kerberos with Active Directory

AUTHORIZATION

Role-Based Access Control
Apache Ranger based Access Control

DATA PROTECTION

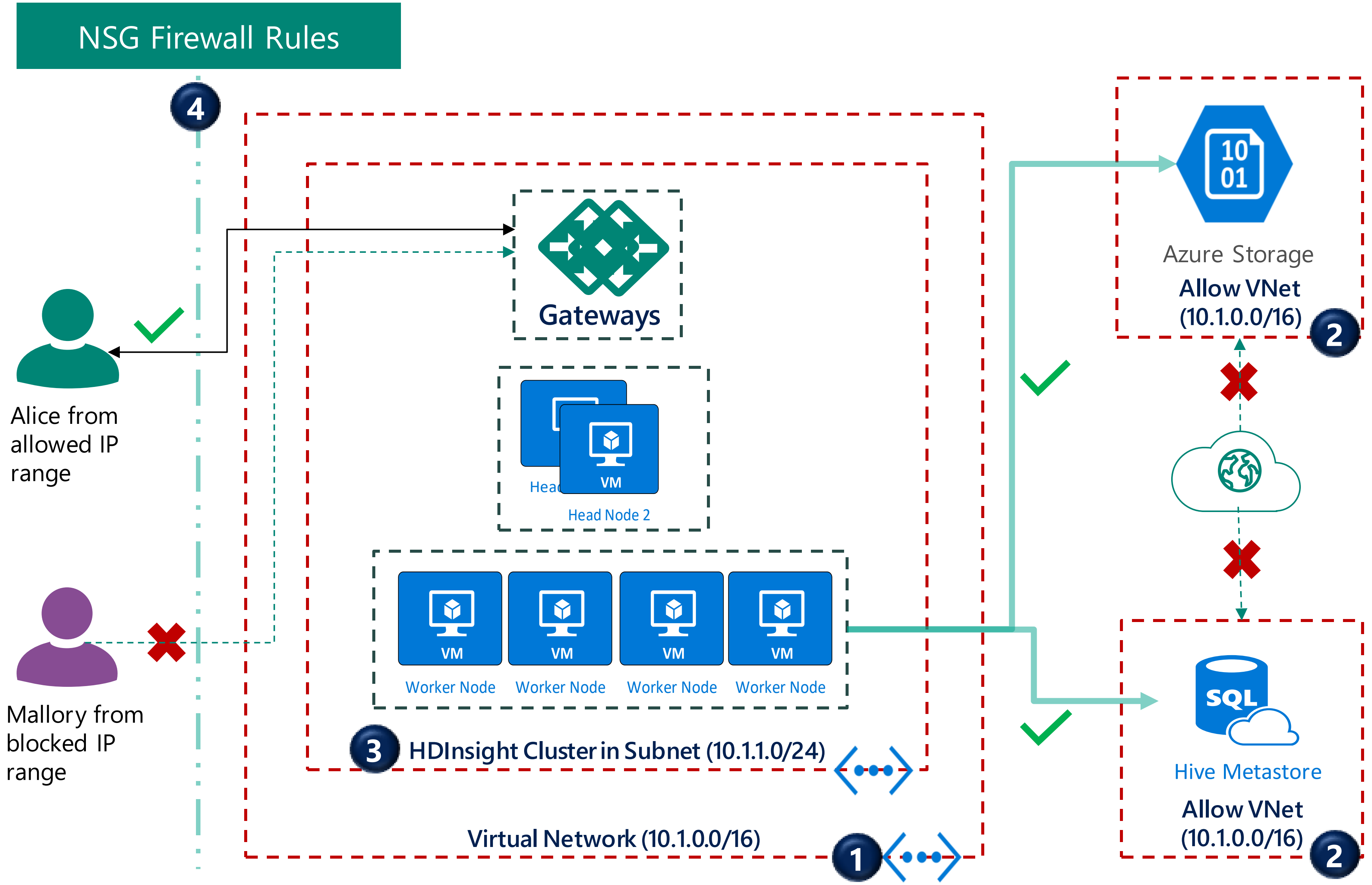
Encryption on-the-wire with HTTPS enforced
Encryption at Rest using Azure Key Vault

Auditing of all data operations and configuration changes

@ashishth

Azure HDInsight Network Security

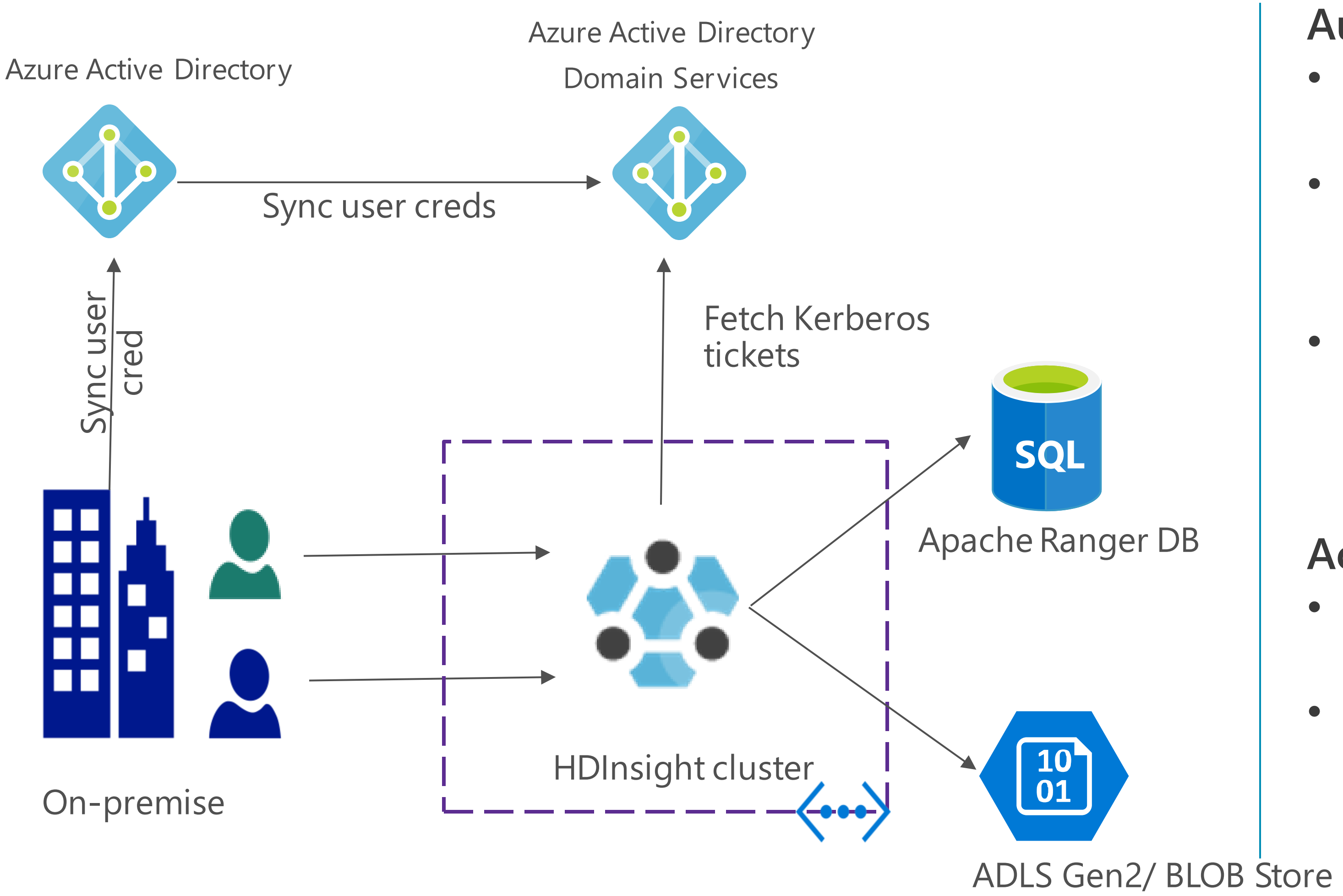
Securing Data sources with Virtual Network Service Endpoints



1. Create VNet, a subnet and enable service endpoint
2. Restrict network access to Storage & SQL
3. Create HDInsight cluster within subnet
4. Create NSG rules to control inbound access to HDInsight cluster

@ashishth

Azure HDInsight: Authentication & Access Control



Authentication:

- Supports identities managed in **Azure Active Directory (AAD)**
- Clusters are joined to **Active Directory Domain Services (ADDS**) based Kerberos Domain Controllers.
- On-premise corporate identities are synced to AAD and ADDS via AD Federation Services.

Access Control:

- Apache Ranger based access control and auditing
- Ranger plugins for Hive, Spark, Kafka and HBase.

@ashishth

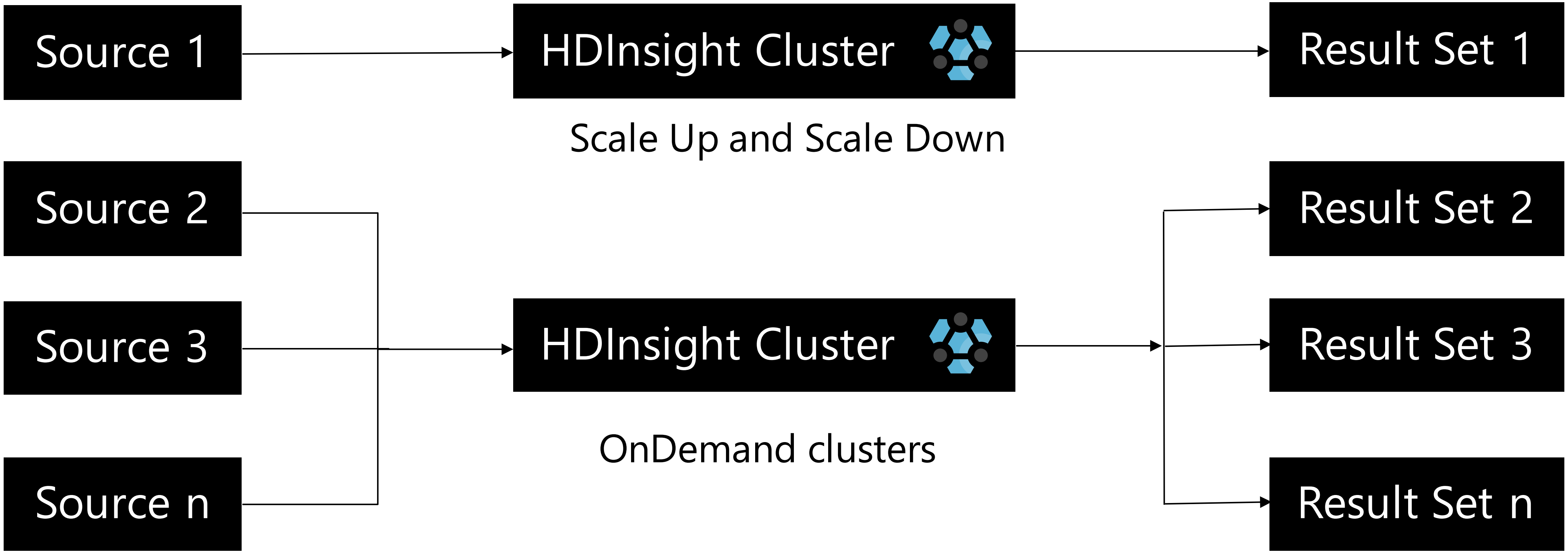
Ranger + ADLS Gen 2 Auth Scenarios in HDInsight

Scenario	Authorizing Component
Yarn: Submit-App	Apache Ranger: Yarn Plugin
Hive Operations: Select , Drop, index, Lock, Read, Write, Masking, Row level filter on Hive Database, Table & Columns	Apache Ranger: Hive Plugin
Create/ Alter Table with storage location reference	Apache Ranger + ADLS Gen 2 ACL's
Spark SQL access with Hive Metastore	Apache Ranger: Hive Plugin
HBase Access Policies	Apache Ranger/ HBase plugin
Kafka Access Policies	Apache ranger/ Kafka Plugin
Access Azure Data Lake Storage Gen2 using the Spark DataFrame API	ADLS Gen 2 ACLs
Access Azure Data Lake Storage Gen2 using the RDD API	ADLS Gen 2 ACLs
HDFS operations: Mkdir, ls, put, copyFromLocal, get, cat, mv, cp etc	ADLS Gen 2 ACLs
Running Map Reduce jobs	ADLS Gen 2 ACLs

@ashishth

Resiliency: The power of embracing failures

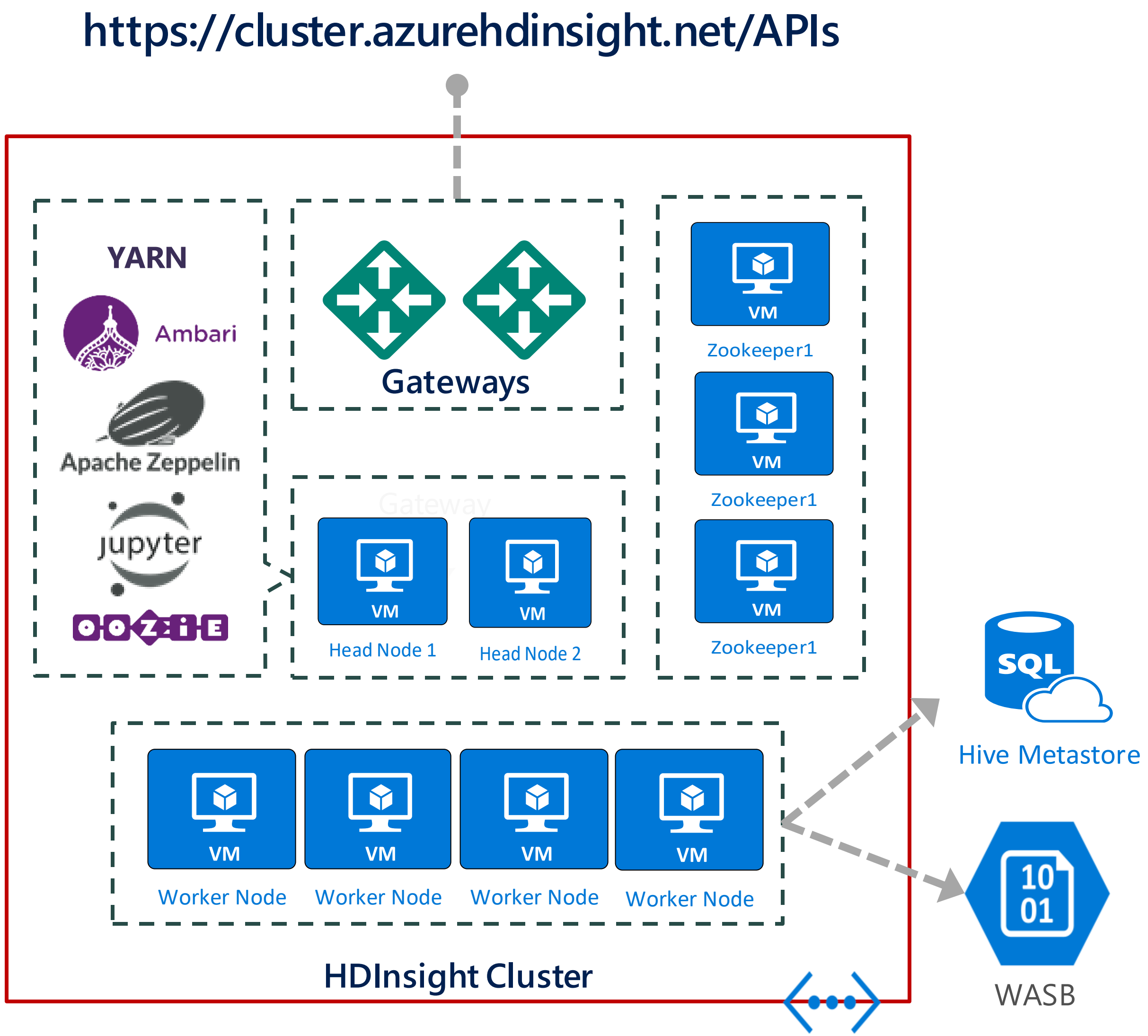
What can go wrong?



Failures of cluster create and scaling operations

@ashishth

Azure HDInsight: Highly Available End-points



Highly Available APIs:

Livy – Spark job submission and interactive session management

Yarn – cluster resource management, Yarn job submission

Ambari – cluster management

Oozie – Oozie workflow scheduling and coordination (legacy APIs, ADF is recommended as a replacement)

<u>Catastrophic Failures and Disasters</u>

@ashishth

What they did?

1. Implemented retry logic for cluster create and scale operations
2. Additional measures for scale down:

Drastic scale down of cluster can get into name node in safe mode

```
hdfs dfsadmin -D 'fs.default.name=hdfs://mycluster/' -safemode get # A report that
shows the

details of HDFS state: hdfs dfsadmin -D 'fs.default.name=hdfs://mycluster/' -report #
Get HDFS

out of safe mode hdfs dfsadmin -D 'fs.default.name=hdfs://mycluster/' -safemode
leave # Get

HDFS into safe mode hdfs dfsadmin -D 'fs.default.name=hdfs://mycluster/' -safemode
enter
```

DR options by workloads

Workload	DR Option
Spark / Hive	Manual, Partner solution
HBase	HBase replication, Snapshot export, Import Export, Copy Tables
Kafka	Mirror Maker

HA & DR

DR options by workloads

Workload	DR Option
Spark / Hive	Manual, Partner solution
HBase	HBase replication, Snapshot export, Import Export, Copy Tables
Kafka	Mirror Maker

More Resources

Spark/ Hive HA & DR https://github.com/anagha-microsoft/hdi-spark-dr

Kafka HA & DR https://github.com/anagha-microsoft/hdi-kafka-dr

HBase Backup, Replication https://docs.microsoft.com/en-us/azure/hdinsight/hbase/apache-hbase-backup-replication

Monitoring

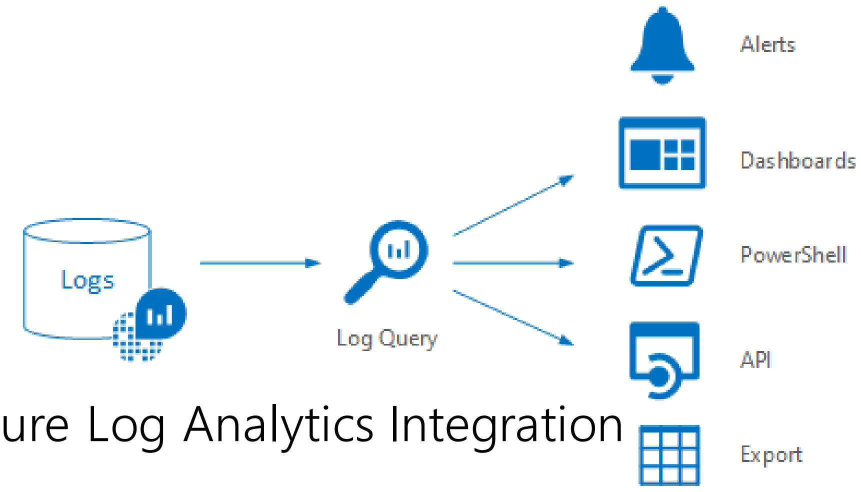
HDInsight Monitoring options

Apache Ambari

- View cluster metrics like CPU, memory, and disk usage at a glance in real time
- Identify malfunctioning components with Ambari alerts
- Monitor queue capacities, jobs, and view associated OSS logs

HDInsight Cluster Metrics

- See gateway requests to monitor cluster stress and cluster size to monitor costs
- Apply filters and chart splitting to extract important data
- Set up alert rules to receive notifications and trigger actions for key metrics



Azure Log Analytics Integration

- Organizes cluster metrics and OSS log records into queryable tables
- Create custom dashboards to surface all the metrics you need from multiple clusters on a single pane of glass

Microsoft

Thank You!

Migrating to Azure HDInsight Guide!

Motivation and benefits covers the benefits of migrating on-premises Hadoop ecosystem components to HDInsight and how to plan for the migration.

Architecture best practices provides best practices for the architecture of HDInsight systems and addresses different types of workloads.

Infrastructure best practices goes into detailed recommendations for managing the infrastructure of HDInsight clusters.

Storage best practices gives recommendations for data storage in HDInsight systems.

Data migration best practices provides recommendations for data migration to HDInsight.

Security and DevOps best practices gives recommendations for security and DevOps in HDInsight systems.

https://azure.microsoft.com/en-us/blog/migrating-on-premises-hadoop-infrastructure-to-azure-hdinsight/