Microsoft Azure

Migrating from Hadoop to the Microsoft Azure Ecosystem

Microsoft Azure is a comprehensive suite of tools and solutions covering areas such as computing, storage, databases, networking, artificial intelligence, machine learning, Internet of Things (IoT), analytics, and more, to cater to various needs in cloud computing. An organization embracing Microsoft Azure can build an efficient cloud-based data platform successfully.

Azure Data Factory

Within this expansive realm, Azure Data Factory emerges as a beacon for Hadoop users, enabling them to seamlessly move data from their storage or processing systems to Azure or between other data sources. With over 11,000 users worldwide, Azure Data Factory ranks 2nd in the Data Integration category (Source: 6sense). As a fully managed, serverless data integration service, Azure Data Factory's code-free intuitive UI allows a single-pane-of-glass for monitoring and management. The platform also supports the migration of SSIS (SQL Server Integration Services) packages to Azure with full compatibility within Azure Data Factory. Additionally, with the SSIS Integration Runtime, users benefit from a fully managed service, eliminating the need for manual infrastructure management.

Azure Synapse Analytics

For the C-suite, Azure Synapse Analytics (formerly known as Azure SQL Data Warehouse) can redefine data management, accelerating time to insight across data warehouses and big data systems. It seamlessly combines SQL technologies from enterprise data warehousing, Spark technologies for large-scale data processing, Data Explorer for log and time series analytics, and pipelines for streamlined data integration and ETL/ELT workflows. This integration extends further with deep collaboration with various Azure services including Power BI, Cosmos DB, and Azure ML.

Azure HDInsight

Azure HDInsight is a managed, full-spectrum, open-source analytics service to simplify running big data frameworks like Hadoop in the Azure environment. Organizations utilize Azure HDInsight in diverse scenarios, such as batch processing (ETL), data warehousing to perform interactive queries at petabyte scales, processing streaming data received in real-time from different sources, and extending existing on-premises Hadoop infrastructure to Azure to leverage the advanced analytics capabilities of the cloud.

Microsoft Fabric

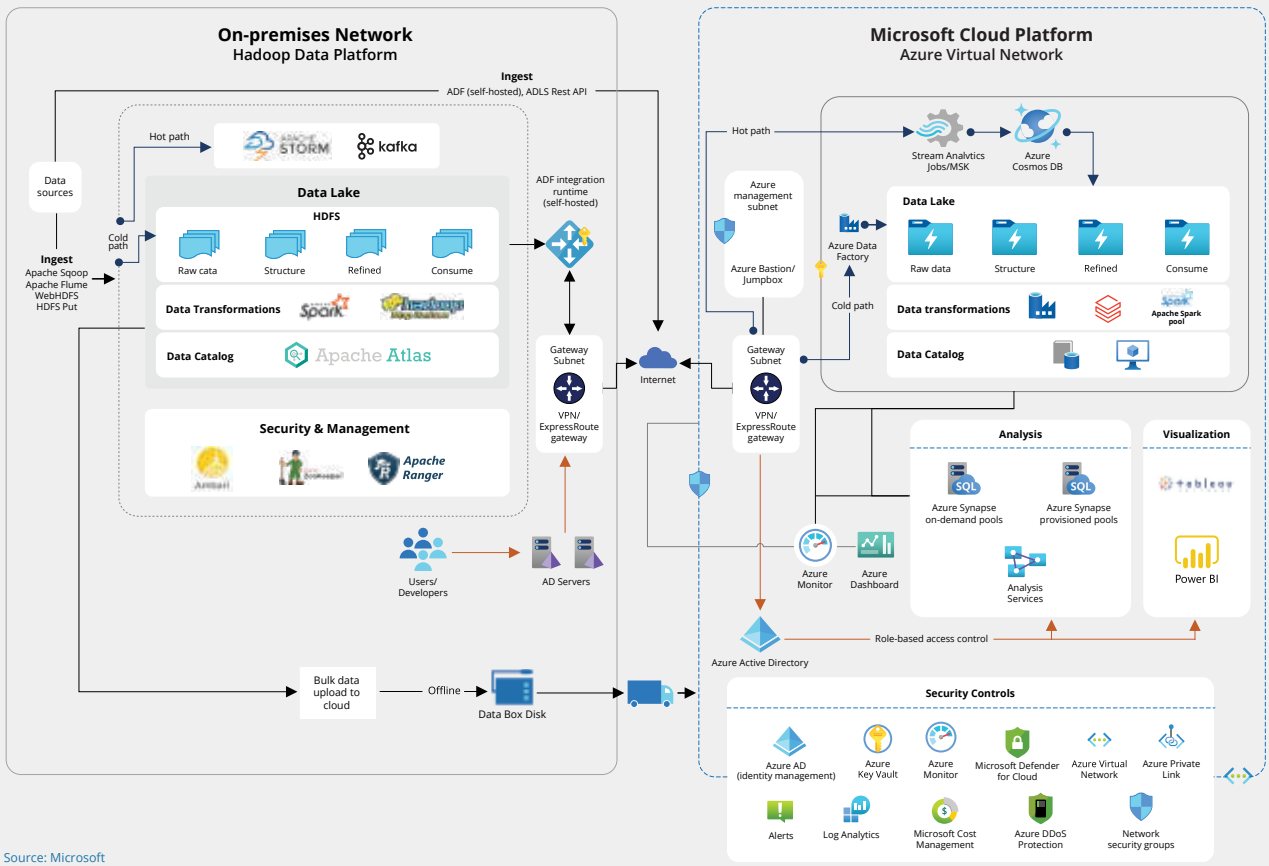
Microsoft Fabric is the brand-new all-in-one analytics solution in this space that is built to address the end-to-end data needs of enterprises. It supports data movement, data science, real-time analytics, and business intelligence by bringing together the latest and existing components from Azure Data Factory, Azure Synapse Analytics, and Power BI onto a shared SaaS foundation. In the age of AI, Microsoft Fabric is infused with Azure OpenAI Services to enable users to harness the potential of generative AI for deeper insights. With this advanced, integrated, and easy-to-use one-stop solution, users can transform their large and complex data repositories into actionable workloads and analytics effortlessly.

Hadoop to Azure Migration Approaches

One of the challenges that organizations face while migrating their workloads from on-premises Hadoop to Azure is getting the deployment right. Organizations contemplating a transition from on-premises Hadoop infrastructure to the expansive landscape of the versatile Microsoft Azure can consider the following three migration approaches to align with the desired end-state architecture and the application.

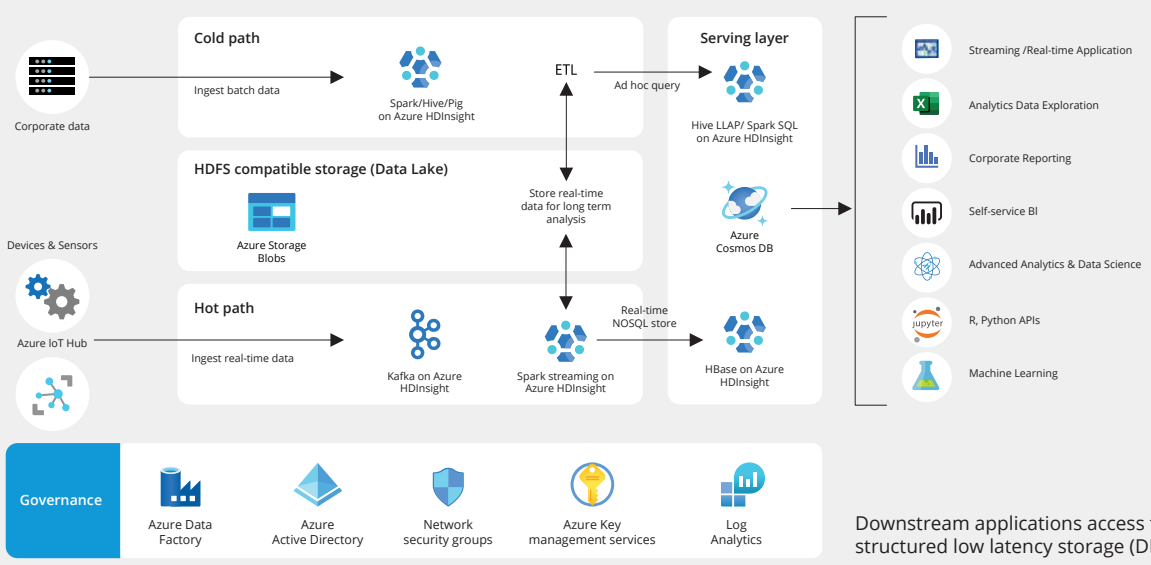
1. **Replatform by Using Azure PaaS Offerings**
2. **Lift and Shift to HDInsight**
3. **Lift and Shift to Azure IaaS**

Replatform by Using Azure PaaS Offerings



Source: Microsoft

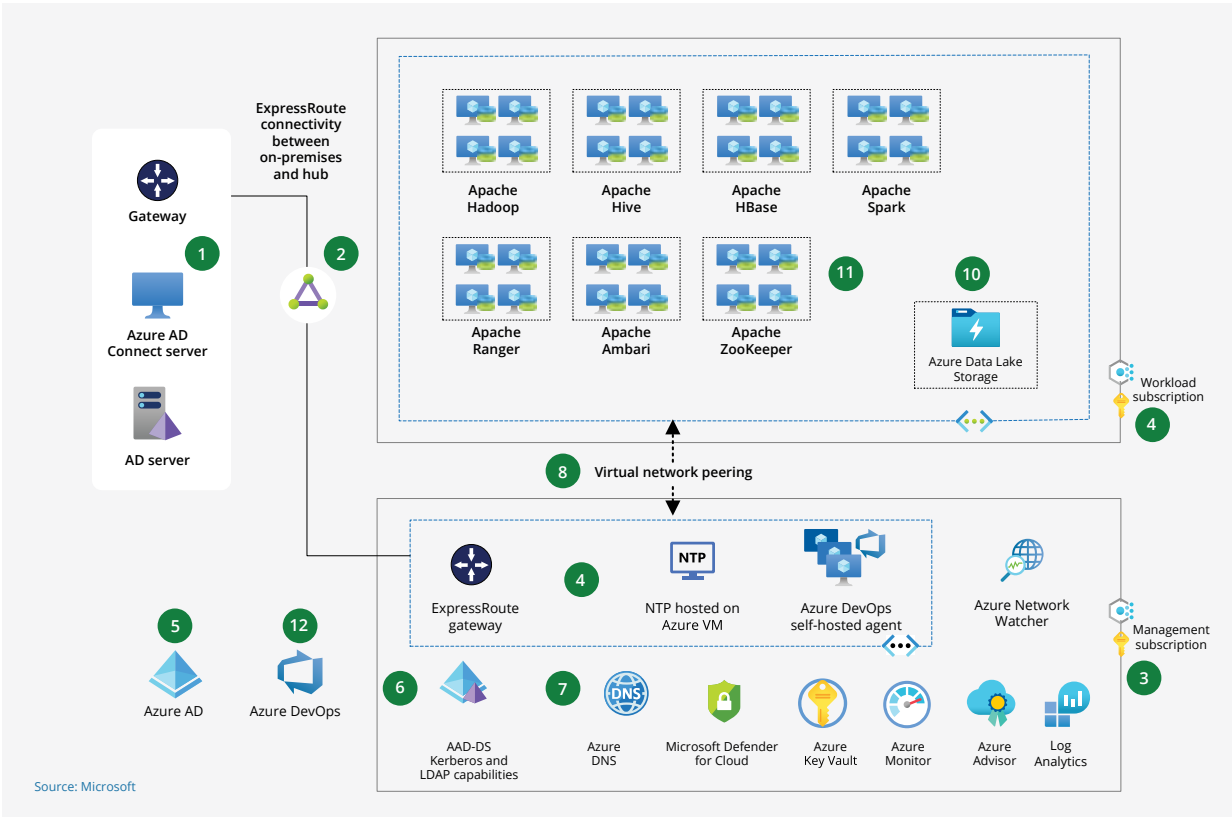
Lift and Shift to HDInsight



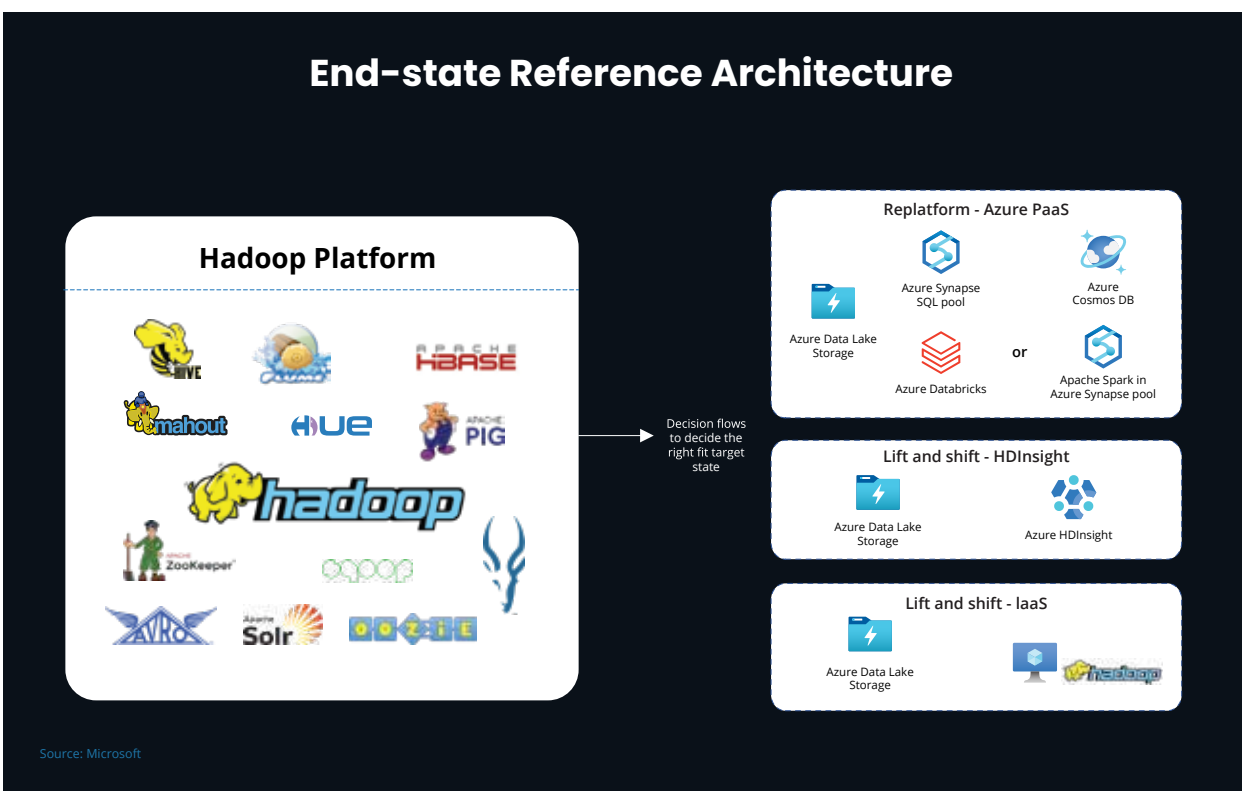
Downstream applications access the structured low latency storage (DBs)

Source: Microsoft

Lift and Shift to Azure IaaS



Source: Microsoft



Source: Microsoft

Migrating From Hadoop to Azure

A Case in Point

How a Major US Bank Derives 3x Times Faster Insights by Shifting to Azure.

Business Challenge

A large bank in the US relying on an on-premises Hadoop platform, faced challenges in efficiently managing the huge influx of data and deriving insights from it. Its Hadoop system was difficult to scale to meet the growing compute demands, query performance was slow, and maintenance overheads were on a constant rise. These challenges affected their ability to promptly access the necessary data for timely insights, thereby limiting democratization opportunities and impacting decision-making processes.

Solution

The bank's technology partner helped them migrate their legacy on-premises Hadoop environment to the Azure ecosystem. The Hadoop workloads were seamlessly migrated to Azure Synapse workloads without any service interruption or data loss.

Value Gained

The fully integrated Azure PaaS platform reduced TCO by up to 50% through its pay-for-what-you-use pricing model and delivered 3x faster data insights.

Mapping Hadoop Components to Azure Services

Hadoop Component	Description	Targeted Azure Services
Apache HDFS	Distributed file system	Azure Data Lake Storage
Apache HBase	Column-oriented table service	HBase on a virtual machine (VM), HBase in Azure HDInsight, Azure Cosmos DB
Apache Spark	Data processing framework	Spark in HDInsight, Azure Synapse Analytics, Azure Databricks
Apache Hive	Data warehouse infrastructure	Hive on a VM, Hive in HDInsight, Azure Synapse Analytics
Apache Ranger	Framework for monitoring and managing data security	Enterprise Security Package for HDInsight, Microsoft Entra ID, Ranger on a VM
Apache Sentry	Framework for monitoring and managing data security	Sentry and Ranger on a VM, Enterprise Security Package for HDInsight, Microsoft Entra ID
Apache MapReduce	Distributed computation framework	MapReduce, Spark
Apache Zookeeper	Distributed coordination service	ZooKeeper on a VM, built-in solution in platform as a service (PaaS)
Apache YARN	Resource manager for the Hadoop ecosystem	YARN on a VM, built-in solution in PaaS
Apache Sqoop	Command line interface tool for transferring data between Apache Hadoop clusters and relational databases	Sqoop on a VM, Sqoop in HDInsight, Azure Data Factory
Apache Kafka	Highly scalable fault-tolerant distributed messaging system	Kafka on a VM, Event Hubs for Kafka, Kafka on HDInsight
Apache Atlas	Open-source framework for data governance and metadata management	Azure Purview

About KANINI

KANINI is a digital transformation enabler, providing cutting-edge software services and solutions that help enterprises drive innovation and business growth. We create impeccable customer experiences through thoughtfully designed digital solutions that help improve our customer's efficiency, scale, and revenues.

We specialize in Cloud Modernization, Data Analytics & AI, Product Engineering, and ServiceNow Consultation and Implementation—all delivered through flexible engagement models.

We focus on empowering Banking and Financial Services, Healthcare, and Manufacturing, among other industries to harness the power of cloud technologies and solutions by implementing agile development practices and a global delivery framework. Find more about our Data Analytics & AI solutions and Consulting services here: https://kanini.com/data-analytics-consulting/

USA | Colombia | Ukraine | UAE | Singapore | India
Contact us at transformations@kanini.com | 615-465-8287