Hadoop migration to Azure

Article • 11/08/2023

Apache Hadoop provides a distributed file system and a framework for using MapReduce techniques to analyze and transform very large data sets. An important characteristic of Hadoop is the partitioning of data and computation across many (thousands) of hosts. Computations are done in parallel close to the data. A Hadoop cluster scales computation capacity, storage capacity, and I/O bandwidth simply by adding commodity hardware.

This article is an overview of migrating Hadoop to Azure. The other articles in this section provide migration guidance for specific Hadoop components. They are:

- Apache HDFS migration to Azure
- Apache HBase migration to Azure
- Apache Kafka migration to Azure
- Apache Sqoop migration to Azure

Hadoop provides an extensive ecosystem of services and frameworks. These articles don't describe the Hadoop components and Azure implementations of them in detail. Instead, they provide high-level guidance and considerations to serve as a starting point for you to migrate your on-premises and cloud Hadoop applications to Azure.

Apache ®, Apache Spark®, Apache Hadoop®, Apache HBase, Apache Hive, Apache Ranger®, Apache Sentry®, Apache ZooKeeper®, Apache Storm®, Apache Sqoop®, Apache Flink®, Apache Kafka®, and the flame logo are either registered trademarks or trademarks of the Apache Software Foundation in the United States and/or other countries. No endorsement by The Apache Software Foundation is implied by the use of these marks.

Hadoop components

The key components of a Hadoop system are listed in the following table. For each component there's a brief description, and migration information such as:

- Links to decision flowcharts for deciding on migration strategies
- A list of possible Azure target services

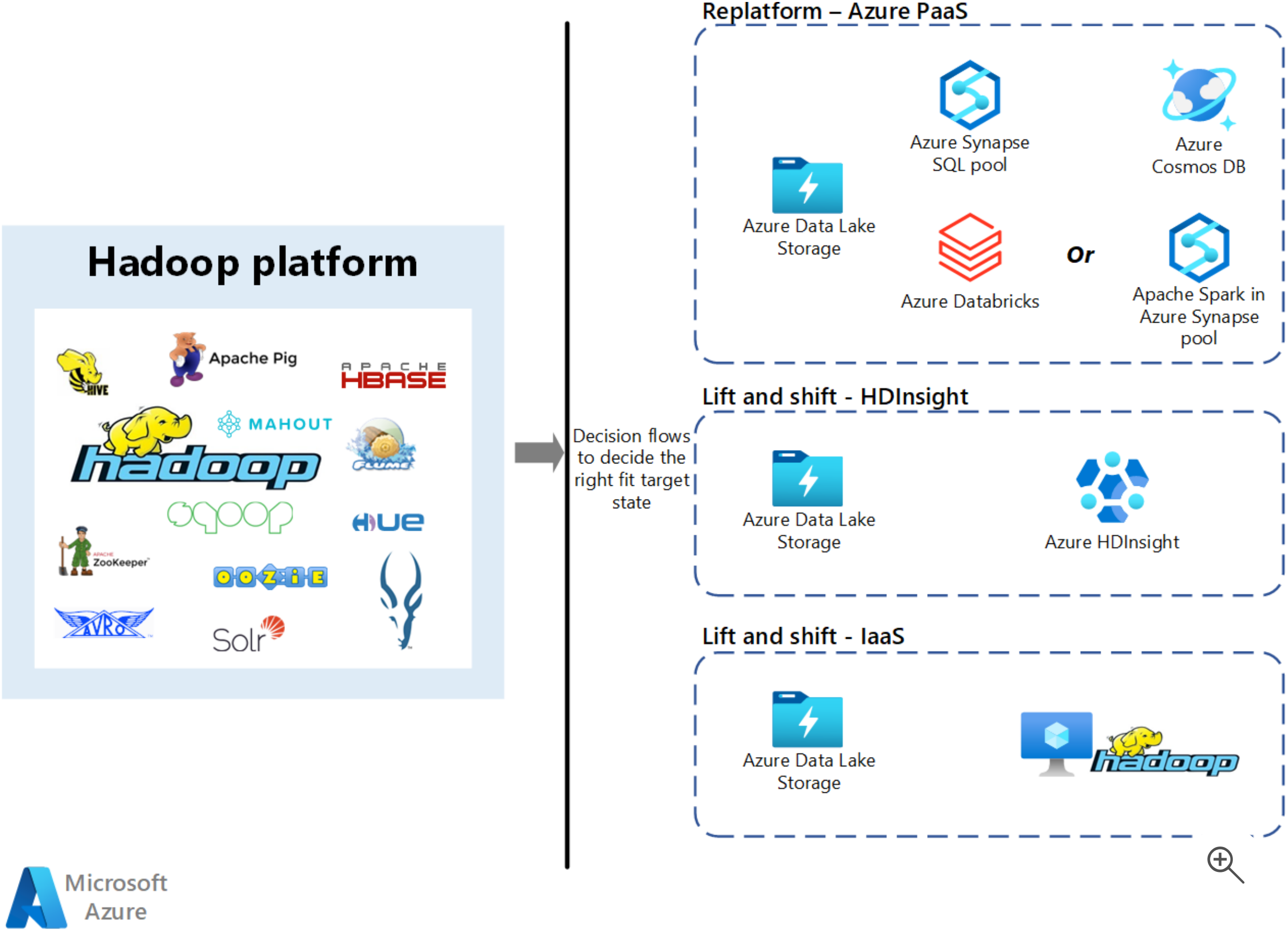
Expand table

Component	Description	Decision flowcharts	Targeted Azure services
Apache HDFS	Distributed file system	Planning the data migration, Pre-checks prior to data migration	Azure Data Lake Storage

Apache HBase	Column-oriented table service	Choosing landing target for Apache HBase, Choosing storage for Apache HBase on Azure	HBase on a virtual machine (VM), HBase in Azure HDInsight, Azure Cosmos DB
Apache Spark	Data processing framework	Choosing landing target for Apache Spark on Azure	Spark in HDInsight, Azure Synapse Analytics, Azure Databricks
Apache Hive	Data warehouse infrastructure	Choosing landing target for Hive, Selecting target DB for Hive metadata	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	Choosing landing targets for Apache Sentry on Azure	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 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	Choosing landing targets for Apache Sqoop on Azure	Sqoop on a VM, Sqoop in HDInsight, Azure Data Factory
Apache Kafka	Highly scalable fault-tolerant distributed messaging system	Choosing landing targets for Apache Kafka on Azure	Kafka on a VM, Event Hubs for Kafka, Kafka on HDInsight
Apache Atlas	Open source framework for data governance and metadata management		Azure Purview

Migration approaches

The following diagram shows three approaches to migrating Hadoop applications:



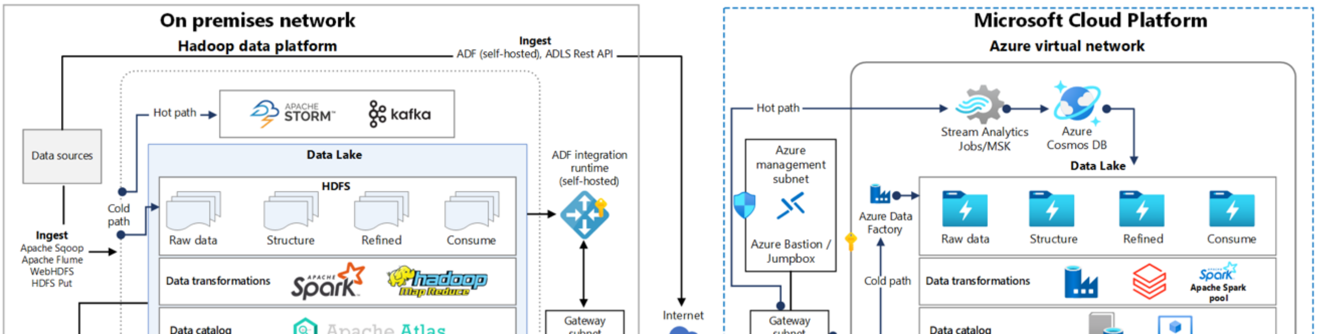
Download a Visio file of this architecture.

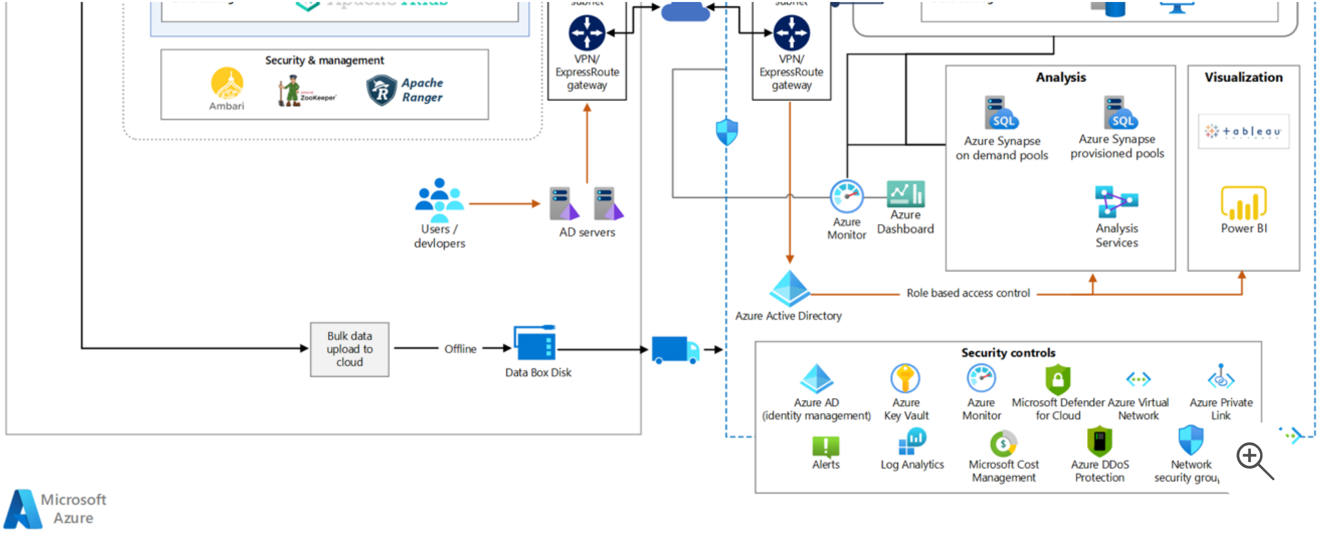
The approaches are:

- **Replatform by using Azure PaaS:** For more information, see Modernize by using Azure Synapse Analytics and Databricks.
- **Lift and shift to HDInsight:** For more information, see Lift and shift to HDInsight.
- **Lift and shift to IaaS:** For more information, see Lift and shift to Azure infrastructure as a service (IaaS).

Modernize by using Azure Synapse Analytics and Databricks

The following diagram shows this approach:

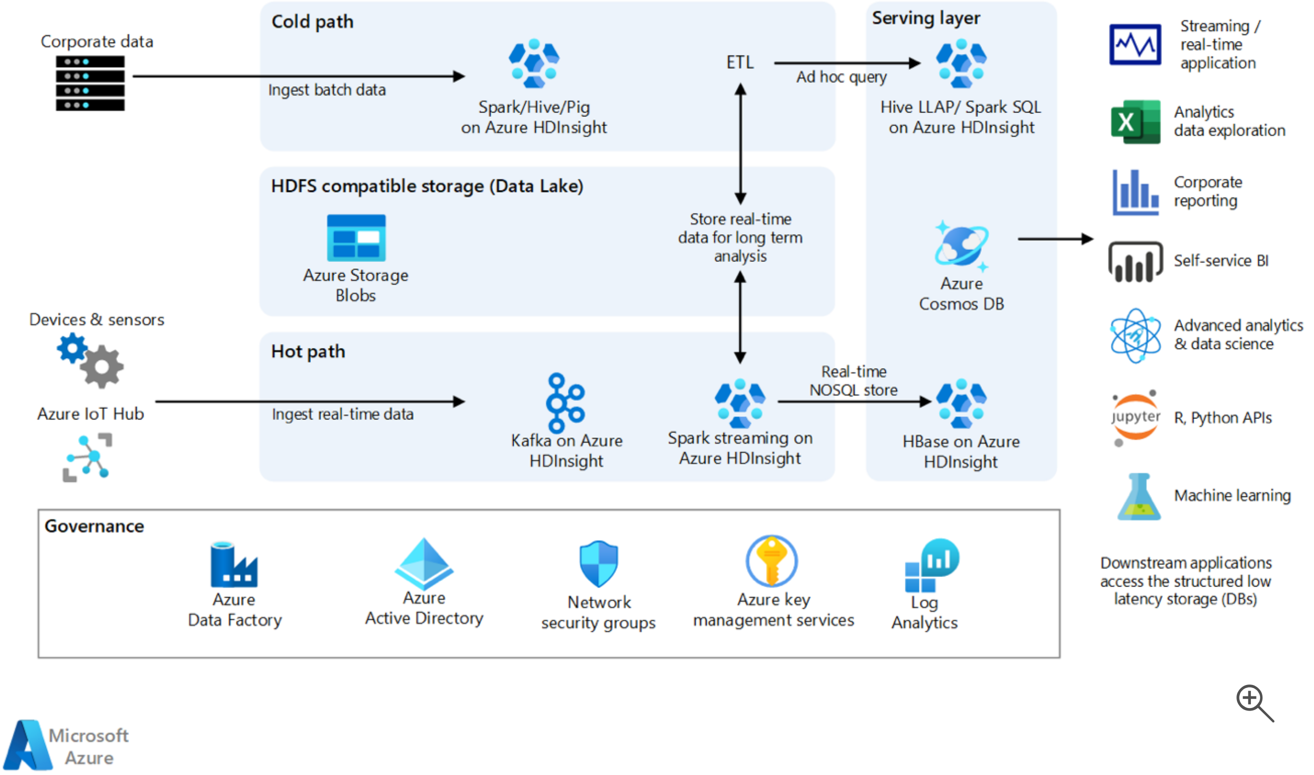




Download a Visio file of this architecture.

Lift and shift to HDInsight

The following diagram shows this approach:

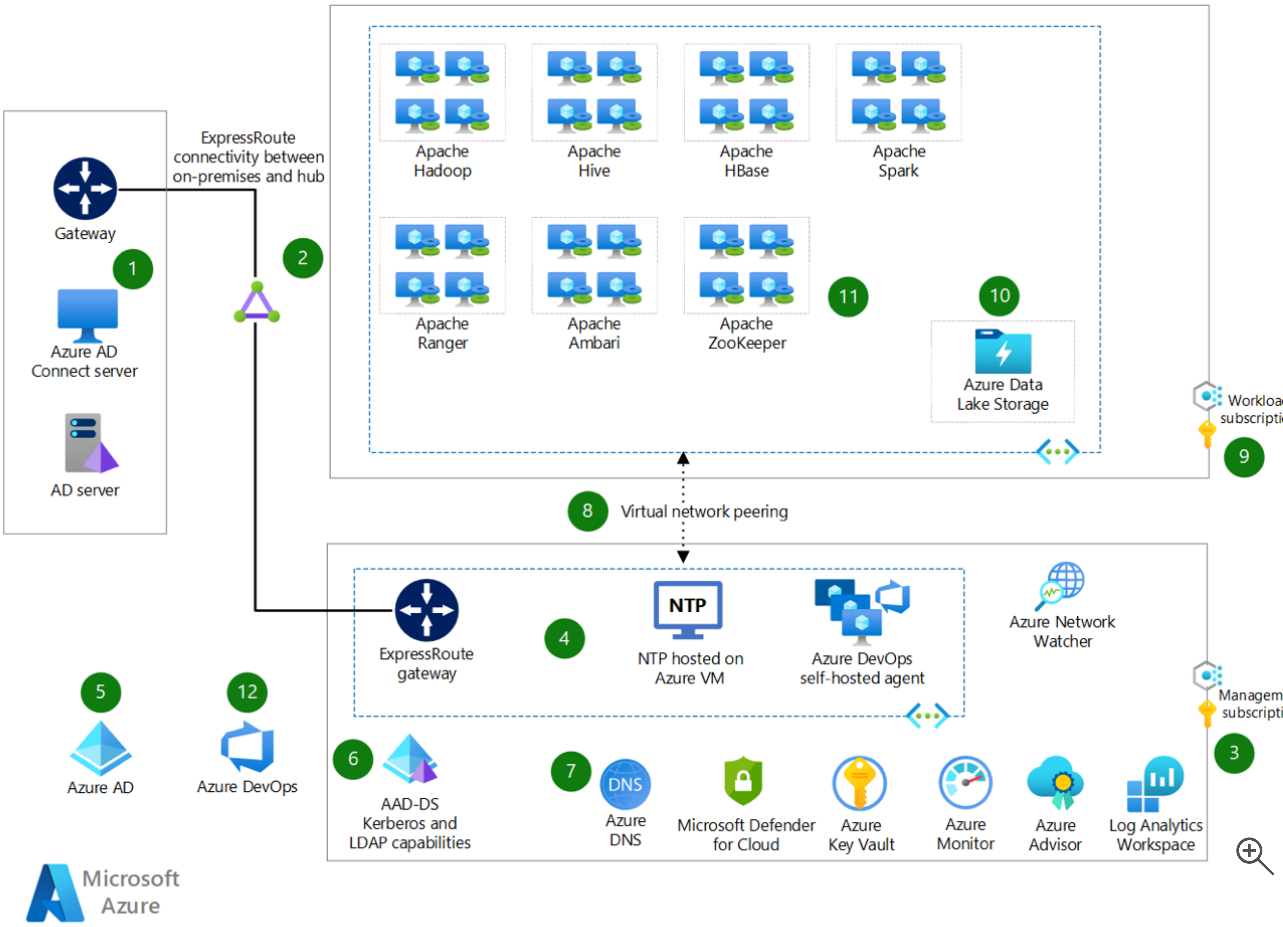


Download a Visio file of this architecture.

For more information, see Guide to migrating Big Data workloads to Azure HDInsight . That article provides a link for downloading a migration guide and also provides an email address that you can use to ask questions or make suggestions.

Lift and shift to Azure infrastructure as a service (IaaS)

The following pattern presents a point of view on how to deploy OSS on Azure IaaS with a tight integration back to on-premises systems such as Active Directory, Domain Controller, and DNS. The deployment follows enterprise-scale landing zone guidance from Microsoft. Management capabilities such as monitoring, security, governance, and networking are hosted within a management subscription. The workloads, all IaaS-based, are hosted in a separate subscription. For more information about enterprise-scale landing zones, see What is an Azure landing zone?.



Download a Visio file of this architecture.

1. On-premises Active Directory synchronizes with Microsoft Entra ID by using Microsoft Entra Connect hosted on-premises.
2. Azure ExpressRoute provides secure and private network connectivity between on-premises and Azure.
3. The management (or hub) subscription provides networking and management capabilities for the deployment. This pattern is in line with enterprise-scale landing zone guidance from Microsoft.
4. The services hosted inside the hub subscription provide network connectivity and management capabilities.

    - **NTP (hosted on Azure VM)** is required to keep the clocks synchronized across all virtual machines. When you run multiple applications, such as HBase and ZooKeeper, you should run a Network Time Protocol (NTP) service or another time-synchronization mechanism on your cluster. All nodes should use the same service for time synchronization. For instructions on setting up NTP on Linux, see 14.6. Basic NTP configuration .

- Azure Network Watcher provides tools to monitor, diagnose, and manage resources in an Azure virtual network. Network Watcher is designed to monitor and repair the network health of IaaS products, including VMs, virtual networks, application gateways, and load balancers.
- Azure Advisor analyzes your resource configuration and usage telemetry and then recommends solutions to improve the cost effectiveness, performance, reliability, and security of your Azure resources.
- Azure Monitor provides a comprehensive solution for collecting, analyzing, and acting on telemetry from your cloud and on-premises environments. It helps you understand how your applications are performing so that you can proactively identify issues that affect the applications and the resources they depend on.
- Log Analytics Workspace is a unique environment for Azure Monitor log data. Each workspace has its own data repository and configuration. Data sources and solutions are configured to store their data in a particular workspace. You need a Log Analytics workspace if you intend to collect data from the following sources:
  - Azure resources in your subscription
  - On-premises computers that are monitored by System Center Operations Manager
  - Device collections from System Center Configuration Manager
  - Diagnostics or log data from Azure Storage
- Azure DevOps Self-Hosted Agent hosted on Azure virtual Machine Scale Sets gives you flexibility over the size and the image of machines on which agents run. You specify a virtual machine scale set, a number of agents to keep on standby, a maximum number of virtual machines in the scale set. Azure Pipelines manages the scaling of your agents for you.

5. The Microsoft Entra ID tenant is synchronized with the on-premises Active Directory via Microsoft Entra Connect synchronization services. For more information, see Microsoft Entra Connect Sync: Understand and customize synchronization.
6. Microsoft Entra Domain Services (Microsoft Entra Domain Services) provides LDAP and Kerberos capabilities on Azure. When you first deploy Microsoft Entra Domain Services, an automatic one-way synchronization is configured and started in order to replicate the objects from Microsoft Entra ID. This one-way synchronization continues to run in the background to keep the Microsoft Entra Domain Services managed domain up-to-date with any changes from Microsoft Entra ID. No synchronization occurs from Microsoft Entra Domain Services back to Microsoft Entra ID.
7. Services such as Azure DNS, Microsoft Defender for Cloud, and Azure Key Vault sit inside the management subscription and provide service/IP address resolution, unified infrastructure security management, and certificate and key management capabilities, respectively.
8. Virtual Network Peering provides connectivity between virtual networks deployed in two subscriptions: management (hub) and workload (spoke).
9. In line with enterprise-scale landing zones, workload subscriptions are used for hosting application workloads.

10. **Azure Data Lake Storage** is a set of capabilities that are built on Azure Blob Storage to do big data analytics. In the context of big data workloads, Data Lake Storage can be used as secondary storage for Hadoop. Data written to Data Lake Storage can be consumed by other Azure services that are outside of the Hadoop framework.
11. **Big data workloads** are hosted on a set of independent Azure virtual machines. Refer to guidance for HDFS, HBase, Hive, Ranger, and Spark on Azure IaaS for more information.
12. **Azure DevOps** is a software as a service (SaaS) offering that provides an integrated set of services and tools to manage your software projects, from planning and development through testing and deployment.

End state reference architecture

One of the challenges of migrating workloads from on-premises Hadoop to Azure is deploying to achieve the desired end state architecture and application. The project that's described in Hadoop Migration on Azure PaaS is intended to reduce the significant effort that's usually needed to deploy the PaaS services and the application.

In that project, we look at the end state architecture for big data workloads on Azure and list the components that are used in a Bicep template deployment. With Bicep we deploy only the modules that we need to deploy architecture. We cover the prerequisites for the template and the various methods of deploying the resources on Azure, such as One-click, Azure CLI, GitHub Actions, and Azure DevOps Pipeline.

Contributors

This article is maintained by Microsoft. It was originally written by the following contributors.

Principal authors:

- Namrata Maheshwary | Senior Cloud Solution Architect
- Raja N | Director, Customer Success
- Hideo Takagi | Cloud Solution Architect
- Ram Yerrabotu | Senior Cloud Solution Architect

Other contributors:

- Ram Baskaran | Senior Cloud Solution Architect
- Jason Bouska | Senior Software Engineer
- Eugene Chung | Senior Cloud Solution Architect
- Pawan Hosatti | Senior Cloud Solution Architect - Engineering
- Daman Kaur | Cloud Solution Architect
- Danny Liu | Senior Cloud Solution Architect - Engineering
- Jose Mendez Senior Cloud Solution Architect
- Ben Sadeghi | Senior Specialist

- Ben Sadeghi | Senior Specialist
- Sunil Sattiraju | Senior Cloud Solution Architect
- Amanjeet Singh | Principal Program Manager
- Nagaraj Seeplapudur Venkatesan | Senior Cloud Solution Architect - Engineering

To see non-public LinkedIn profiles, sign in to LinkedIn.

Next steps

Azure product introductions

- Introduction to Azure Data Lake Storage Gen2
- What is Apache Spark in Azure HDInsight?
- What is Apache Hadoop in Azure HDInsight?
- What is Apache HBase in Azure HDInsight?
- What is Apache Kafka in Azure HDInsight?
- Overview of enterprise security in Azure HDInsight

Azure product reference

- Microsoft Entra documentation
- Azure Cosmos DB documentation
- Azure Data Factory documentation
- Azure Databricks documentation
- Azure Event Hubs documentation

- Azure Functions documentation
- Azure HDInsight documentation
- Microsoft Purview data governance documentation
- Azure Stream Analytics documentation
- Azure Synapse Analytics

Other

- Enterprise Security Package for Azure HDInsight
- Develop Java MapReduce programs for Apache Hadoop on HDInsight
- Use Apache Sqoop with Hadoop in HDInsight
- Overview of Apache Spark Streaming
- Structured Streaming tutorial
- Use Azure Event Hubs from Apache Kafka applications

Related resources

- Apache HDFS migration to Azure
- Apache HBase migration to Azure
- Apache Kafka migration to Azure
- Apache Sqoop migration to Azure

Feedback

Was this page helpful? Yes No