Componentes clave

Azure HDInsight

Cluster			
Son grupos de máquinas virtuales que trabajan juntas para ejecutar cargas de trabajo de análisis de datos. Cada cluster incluye varios nodos.			
Nodo maestro (Principal) head nodes	Nodo de trabajo worker nodes	Nodo de borde gateway	Nodos ZooKeeper
Encargado de administrar y coordinar el clúster	Encargado de ejecutar tareas de procesamiento de dato	Proporcionar un punto de acceso externo al clúster	coordinan las operaciones distribuidas entre los nodos
Opciones de escalado			
Automatico		Manual	

Servicios de Datos				
Procesamiento distribuido	**Procesamiento en tiempo real**	**Datos no estructurados**	**Consultas**	**Transmisión**
Descripción	Marco de procesamiento distribuido de código abierto para el almacenamiento y procesamiento de grandes conjuntos de datos.	Marco de procesamiento de datos de código abierto para el procesamiento en tiempo real y análisis de grandes conjuntos de datos	Una base de datos NoSQL de código abierto para el almacenamiento de datos no estructurados y semi-estructurados.	Un marco de procesamiento de datos de código abierto para consultas y análisis de datos en Hadoop	Plataforma de procesamiento de flujo de datos de código abierto para el procesamiento en tiempo real de grandes volúmenes de datos de transmisión.
Se utiliza	Para el procesamiento de datos a gran escala en lotes	Para el procesamiento de datos en tiempo real y análisis de datos	Para consultas ad-hoc y análisis de datos en Hadoop	Para aplicaciones que requieren una alta velocidad de lectura y escritura y acceso aleatorio a datos	Para aplicaciones de transmisión de datos en tiempo real

Almacenamiento	
Nube	**Local**
Azure Blob Storage / Azure Data Lake Storage	Almacenamiento en el disco duro local de los nodos en el clúster de HDInsight
• Datos no estructurados • Precios mas bajos (acceso frecuente a datos) • Escalado de grandes volúmenes de datos • Puede utilizar con varios servicios de procesamiento de datos	• Datos no estructurados y estructurados • Escalado de grandes conjuntos de datos (petabytes de almacenamiento) • Se integra de manera nativa con HDInsight. • Optimizado para trabajar con servicios de procesamiento de datos específicos de Azure. • Mas costoso	HDFS Hadoop Distributed File System
		Datos temporales y los resultados intermedios del procesamiento
Cual Elegir		
Si necesitas almacenar grandes conjuntos de datos y quieres una esca-labilidad y durabilidad alta alta disponibilidad y durabilidad de los datos		Si necesitas acceder a los datos con frecuencia y a alta velocidad más rápido que el almacenamiento en la nube
Si necesitas una combinación de esca-labilidad y velocidad de acceso a los datos, puedes utilizar ambos tipos de almacenamiento en HDInsight. Puedes almacenar los datos principales en el almacenamiento en la nube y utilizar el almacenamiento local en el clúster para almacenar datos temporales y resultados intermedios del procesamiento.		

Herramientas de desarrollo y administración				
Portal de Azure	AZure CLI	Azure Powershell		
Azure Monitor	Azure Data Factory	Power BI	AZure Databricks	
Azure HDInsight .NET SDK	Python	R	Java	Scala
aplicaciones web integradas				
Apache Ambari	Apache Zeppelin	Jupiter Notebook	Hue	

Seguridad	
Azure Key Vault	Azure Active Directory Domain Services (Azure AD DS)
Control de acceso basado en roles (RBAC)	Microsoft Defender for cloud
Azure Storage Service Encryption	Auditoría y registro de actividad