DP-203 Microsoft Azure Data Engineer

Day8 – Azure HDInsight

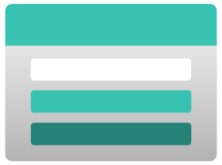
1st Aug 2021

Vinodkumar Bhovi

databoq®

Data

Data Storage



Azure Storage Accounts

Azure Cosmos DB

Azure Data Lake

Azure SQL

Azure Synapse Analytics

Data Transformation

Azure Data Factory

Azure Stream Analytics

Azure Databricks

Azure HDInsight

© 2021 databag.ai – Proprietary and Confidential

databag®

Agenda

- Why Hadoop is Hard?
- Why Hadoop on cloud?
- How HDInsight makes Hadoop easy?
- Important aspects of Hadoop?
- HDInsight Architecture

© 2021 databag.ai – Proprietary and Confidential

databag®

3 Challenges with Hadoop?

© 2021 databag.ai – Proprietary and Confidential

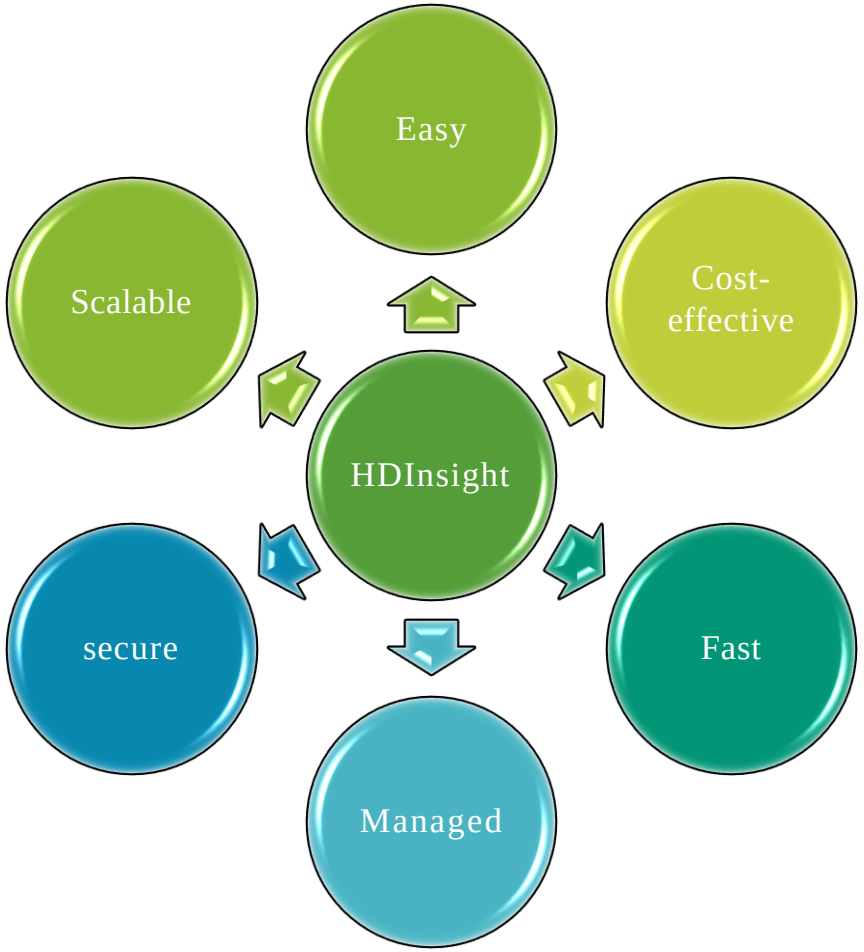
HDInsight makes Hadoop easy

© 2021 databag.ai – Proprietary and Confidential

databag®

What is Azure HDInsight?

HDInsight is a cloud distribution of Hadoop components



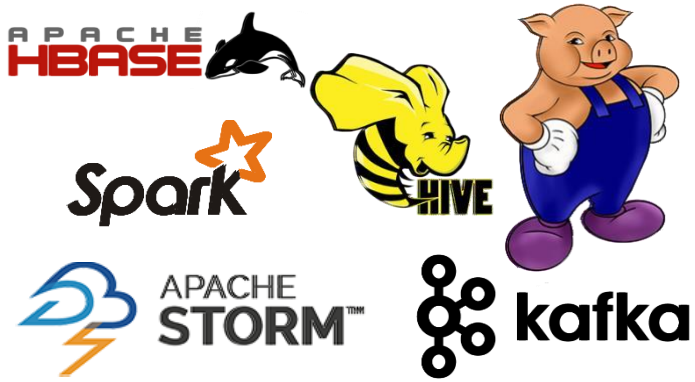
© 2021 databag.ai – Proprietary and Confidential

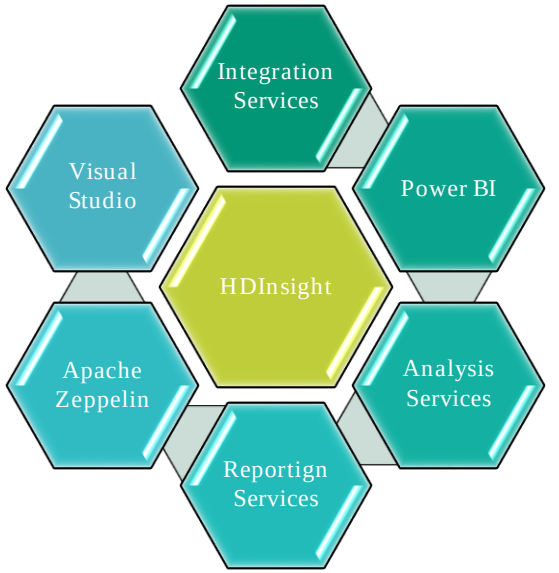
databag®

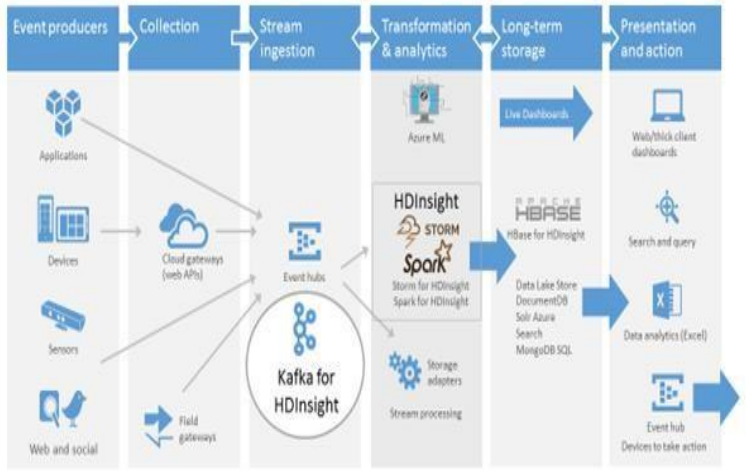
HDInsight makes Hadoop easy

© 2021 databag.ai – Proprietary and Confidential

Important aspects of HDInsight



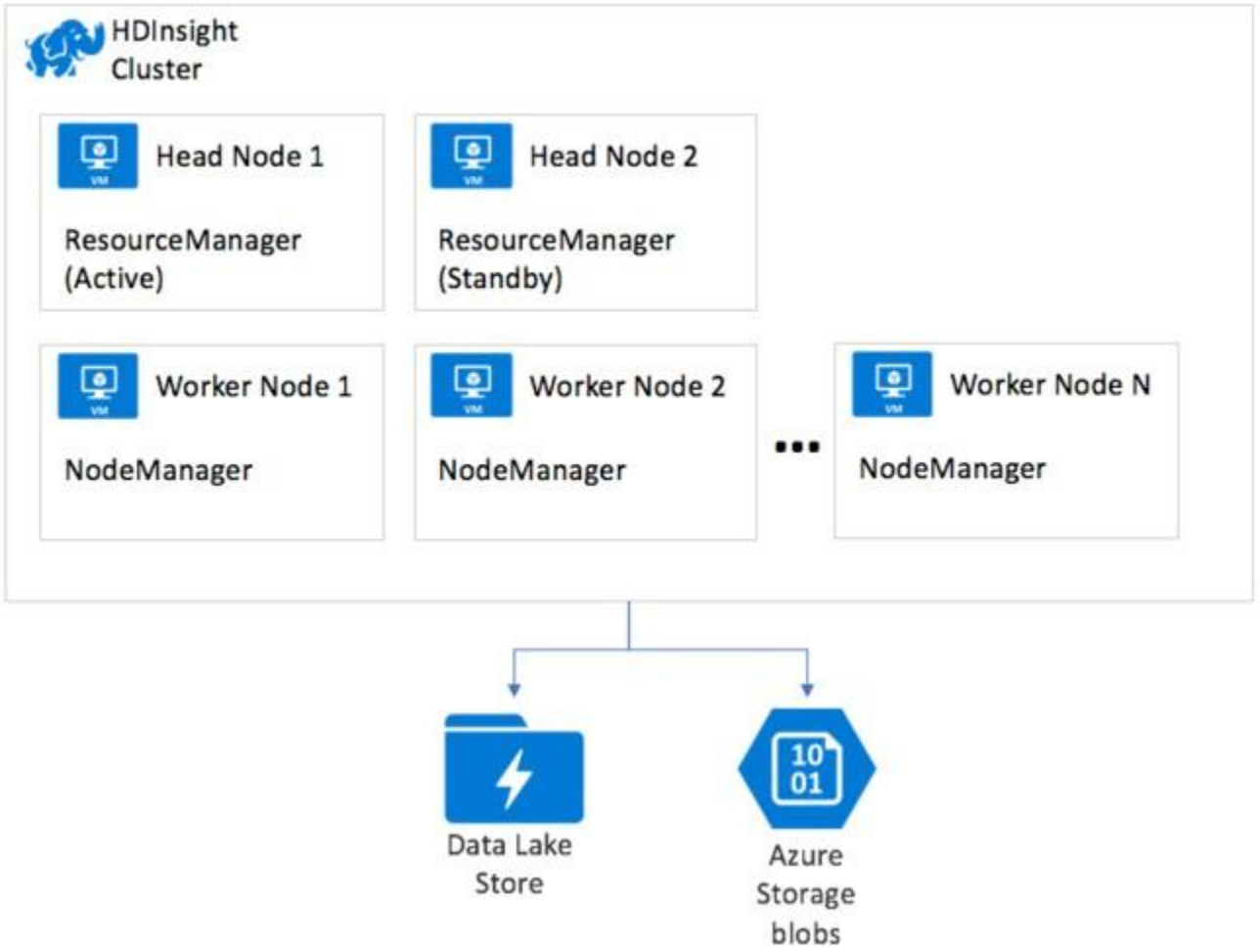




© 2021 databag.ai – Proprietary and Confidential

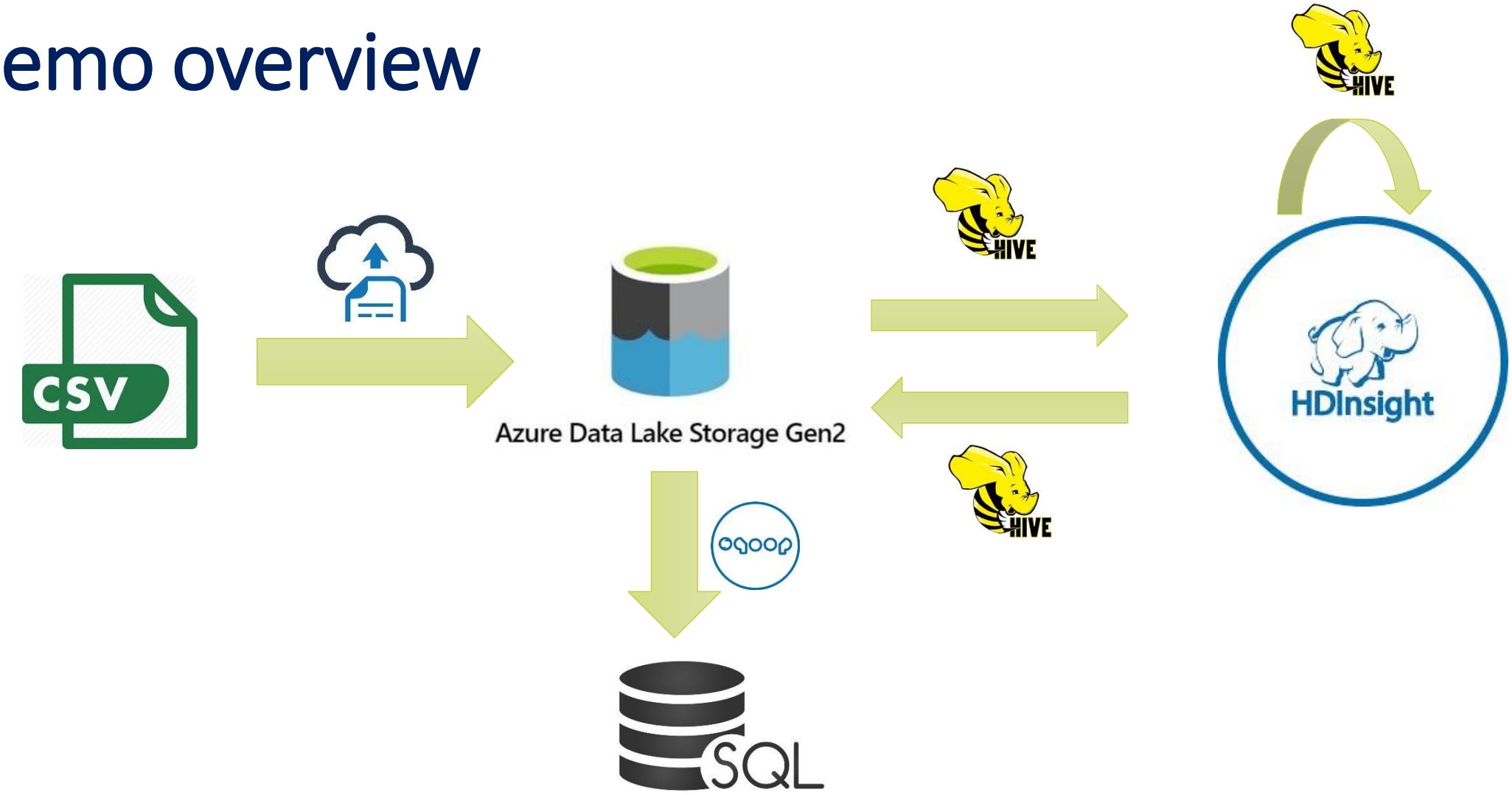
databag®

Hierarchy



© 2021 databag.ai – Proprietary and Confidential

Demo overview



© 2021 databag.ai – Proprietary and Confidential

databag®

Ambari

"Ambari is a Hadoop management platform responsible for cluster administration, monitoring and configuration"

© 2021 databag.ai – Proprietary and Confidential

- Visualized reports
- System monitoring
- Change configuration
- Manage authentication

© 2021 databag.ai – Proprietary and Confidential

databag®

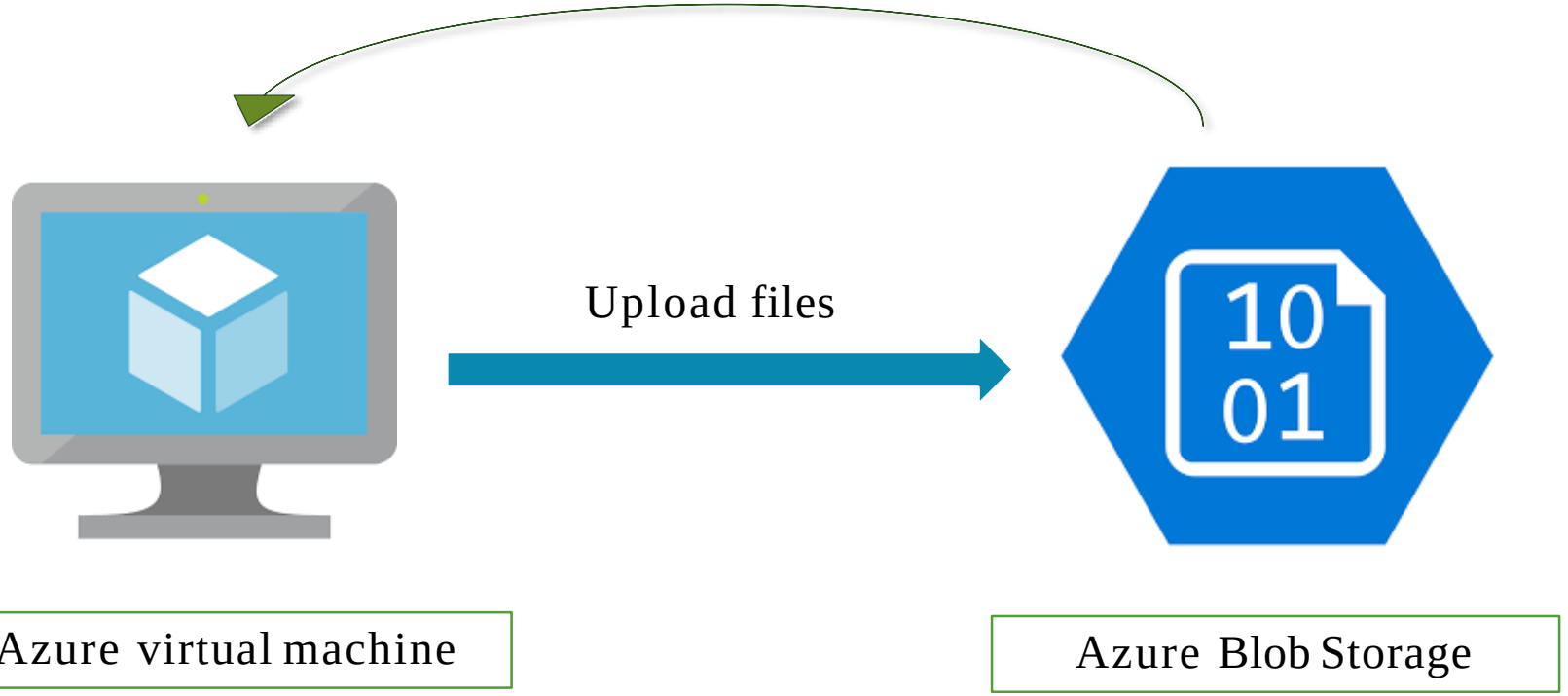
Managed Identity

Managed identities are
used by Azure services to authenticate to other Azure services
that support Azure AD authentication.

© 2021 databag.ai – Proprietary and Confidential

databag®

Managed Identity

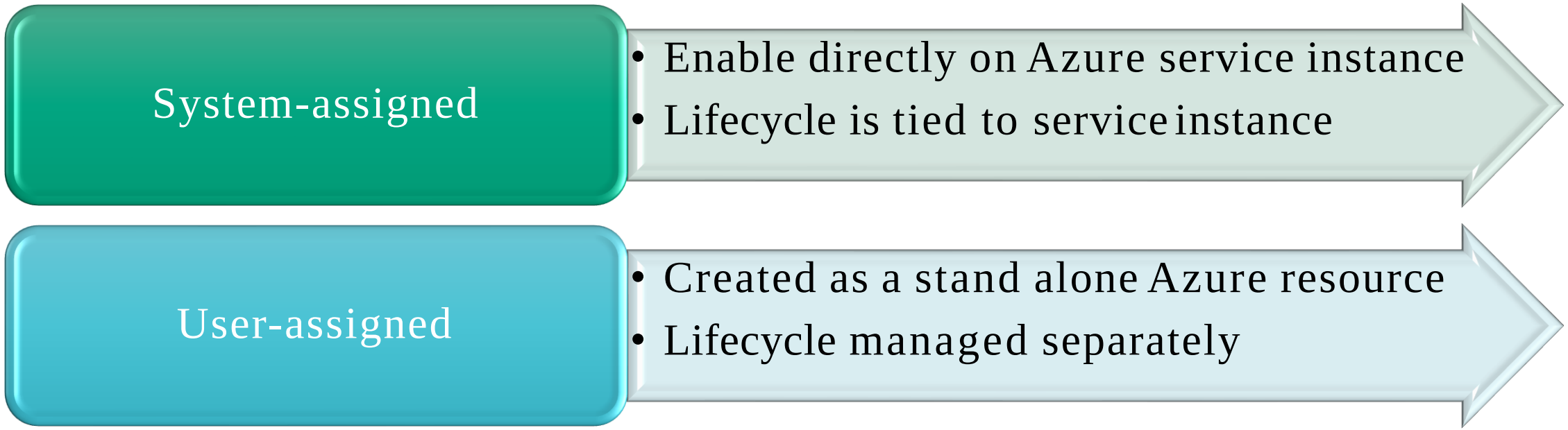


© 2021 databag.ai – Proprietary and Confidential

Managed Identity – 2 steps

© 2021 databag.ai – Proprietary and Confidential

Two types of Managed Identity



© 2021 databag.ai – Proprietary and Confidential

databag®

HDInsight cluster types

- **Hadoop** – Batch query and analysis of HDFS stored data
- **HBase** – Processing for large schema less NoSQL data
- **Interactive Query** – In-memory caching for fast Hive queries
- **Kafka** – Distributed streaming data platform
- **ML Services** – Predictive modeling and machine learning
- **Spark** – In-memory processing and interactive queries
- **Strom** – Real-time event processing

© 2021 databag.ai – Proprietary and Confidential

databag®

Thank you!

© 2021 databag.ai – Proprietary and Confidential

databag®