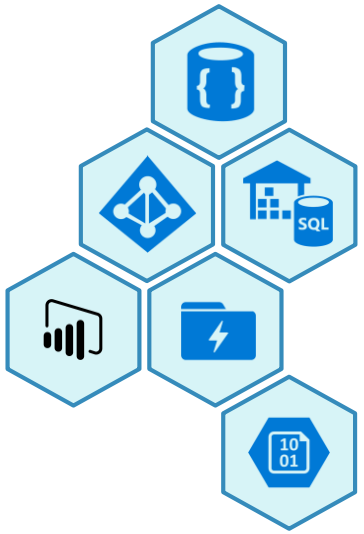
Microsoft Azure HDInsightトレーニング
Step-1: Azure HDInsightの特徴

目次

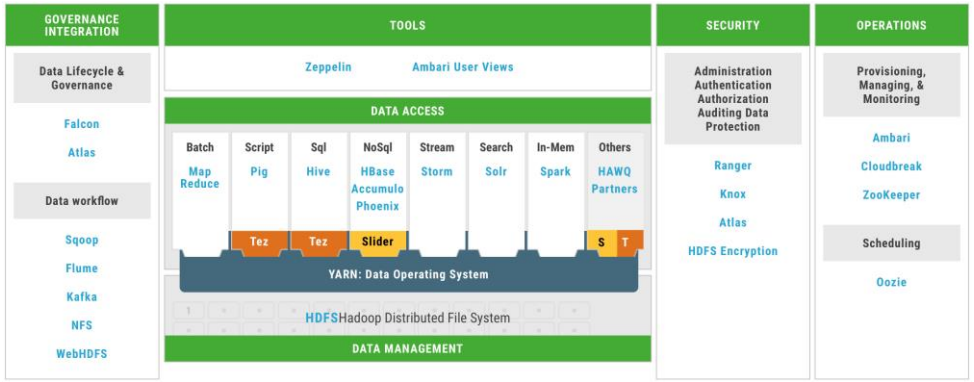
1. Azure HDInsightとは
2. Azure HDInsightの特徴
    1. Azure Storage Blob/Azure Data Lakeとの連携
    2. Apache Sparkによる強力なETL処理
    3. Hive on Tez + PowerBI によるダッシュボード

1.1. Azure HDInsightとは



Microsoft + HORTONWORKS

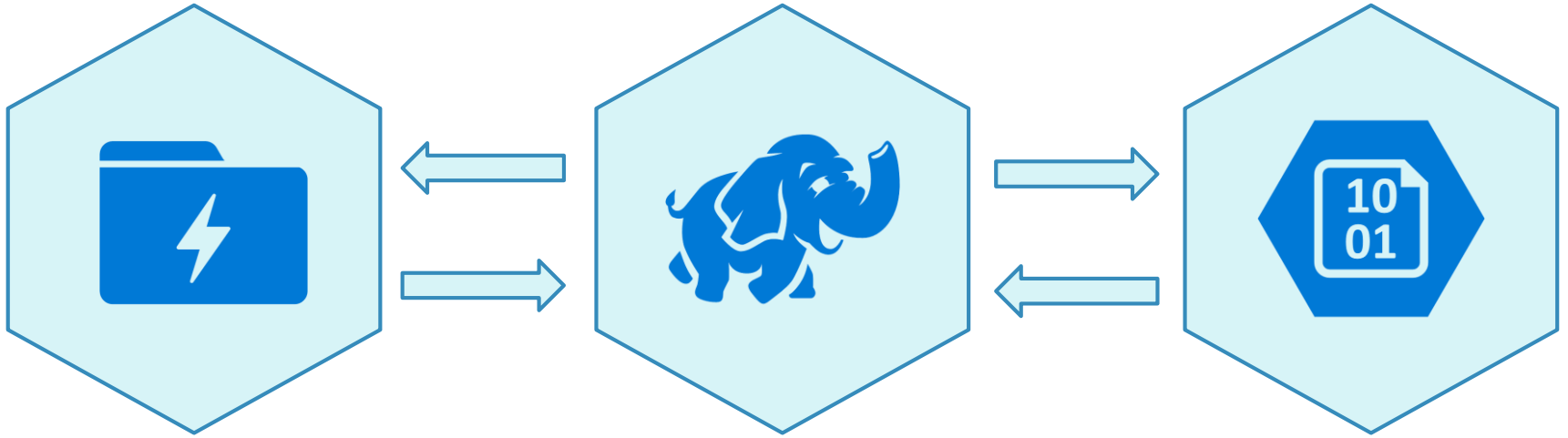
PREMIER DATA CLOUD PROVIDER



- Microsoft Azure のマネージド Hadoop サービス
- HDP(Hortonworks Data Platform)がベース
- Windows版はMicrosoftがHortonworksと協業して開発*1
- Azure の各種サービスと連携可能

*1: https://hortonworks.com/products/cloud/azure-hdinsight/

Azure HDInsightの特徴(1)
Azure Storage Blob/Azure DataLakeとの連携



- BLOB ストレージは大容量・低価格・高信頼性を兼ね備えたMicrosoft Azure の主要サービス
- Azure 上で稼働する多くのアプリケーションが、 BLOB ストレージにデータを保存している
- HDInsightはStorage Blob内のデータに直接処理を実行することが可能
- HDInsight内のHDFSも利用可能

Azure HDInsightの特徴(2)
Apache Sparkによる強力なETL処理

- Apache Sparkの機能SparkSQLで様々なデータソースにSQLライクなクエリをかけて、データを加工(ETL処理)することが可能。ETL処理はWorkerノード全体にまたがり分散実行
- Apache Spark 2.0以降のSparkSQLはSQL:2003準拠。既存のSQLコードの移植性が大幅に向上
- SQL処理にありがちなTEMPORALY TABLE不要。オンメモリ処理+遅延実行により効率的な処理が可能

Azure HDInsightの特徴(3)
Hive on Tez + PowerBIによるダッシュボード

- HDInsightのHive実行エンジンTezは従来のMapReduceエンジンに比べ最大約100倍のSQL処理高速化を実現
- Hiveをバッチ処理だけではなく、ダッシュボードのバックエンドとしても利用できる可能性
(すべてのユースケースが該当するわけではない)
- PowerBIからは複数のドライバ(ODBC, DirectQuery)でHDInsightに接続可能