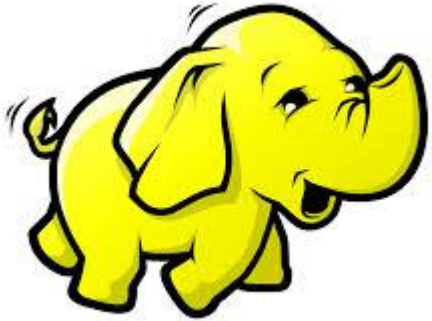
FPAcademy

Sebastian Zęderowski

Azure as a Big Data Platform (HDInsight)

Azure as a Big Data Platform

Sebastian Zęderowski



Usługa Azure HDInsight jest dystrybucją w chmurze składników usługi Hadoop z platformy Hortonworks Data Platform (HDP).

Azure HDInsigh – typy klastra

- **Apache Hadoop**: korzysta z systemu HDFS, zarządzania zasobami YARN i prostego modelu programowania MapReduce do celów równoległego przetwarzania i analizowania danych.

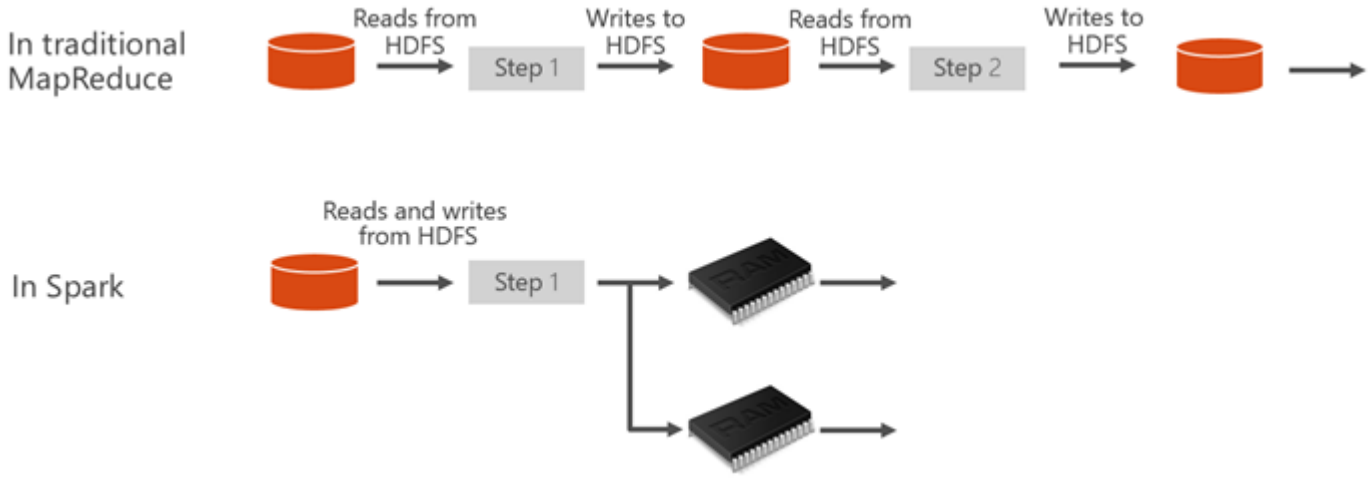
- **Apache Spark:** platforma przetwarzania równoległego obsługująca przetwarzanie w pamięci w celu zwiększania wydajności aplikacji do analizy danych big data. Platforma Spark jest odpowiednia do języka SQL, strumieniowego przesyłania danych oraz uczenia maszynowego.

- **Apache HBase:** baza danych NoSQL oparta na platformie Hadoop, która zapewnia dostęp losowy i wysoki poziom spójności w przypadku dużych ilości danych z częściową strukturą lub bez struktury — potencjalnie miliardów wierszy z milionami kolumn.

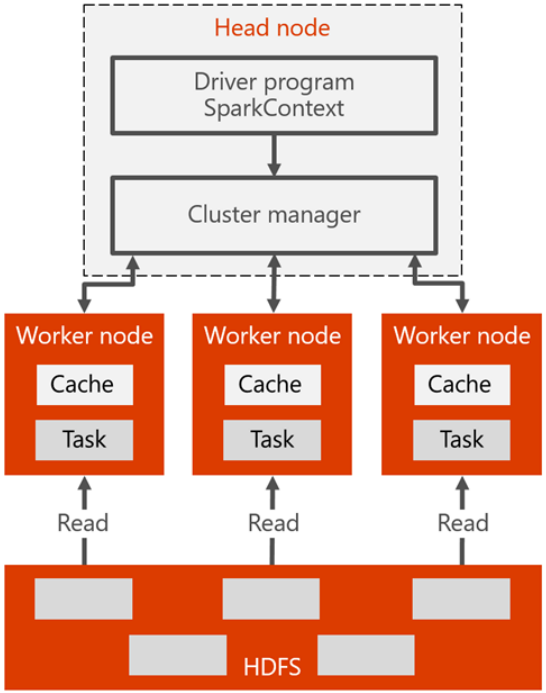
Azure HDInsigh – typy klastra

- **Microsoft R Server:** serwer przeznaczony do hostowania równoległych, rozproszonych procesów języka R oraz zarządzania nimi.
- **Apache Storm:** rozproszony system obliczeniowy działający w czasie rzeczywistym do szybkiego przetwarzania dużych strumieni danych.
- **Apache Hive (LLAP) :** buforowanie w pamięci umożliwiające interaktywne i szybsze zapytania programu Hive.
- **Apache Kafka:** platforma służąca do tworzenia potoków danych przesyłanych strumieniowo i aplikacji do obsługi tych danych. Platforma Kafka obejmuje również funkcję kolejki komunikatów, która umożliwia publikowanie i subskrybowanie strumieni danych.

Spark vs MapReduce



Architektura klastra Spark



Spark

Centralną koncepcją Apache Spark są tzw. *resilient distributed datasets*, co można przetłumaczyć jako odporne na awarie, rozproszone kolekcje danych. Istnieją one na wielu maszynach jednocześnie (na dyskach lub w pamięci operacyjnej), a w przypadku awarii są odtwarzane.

Na tych kolekcjach danych można przeprowadzać określone operacje, czyli – zgodnie ze słownikiem Spark – transformacje i akcje:

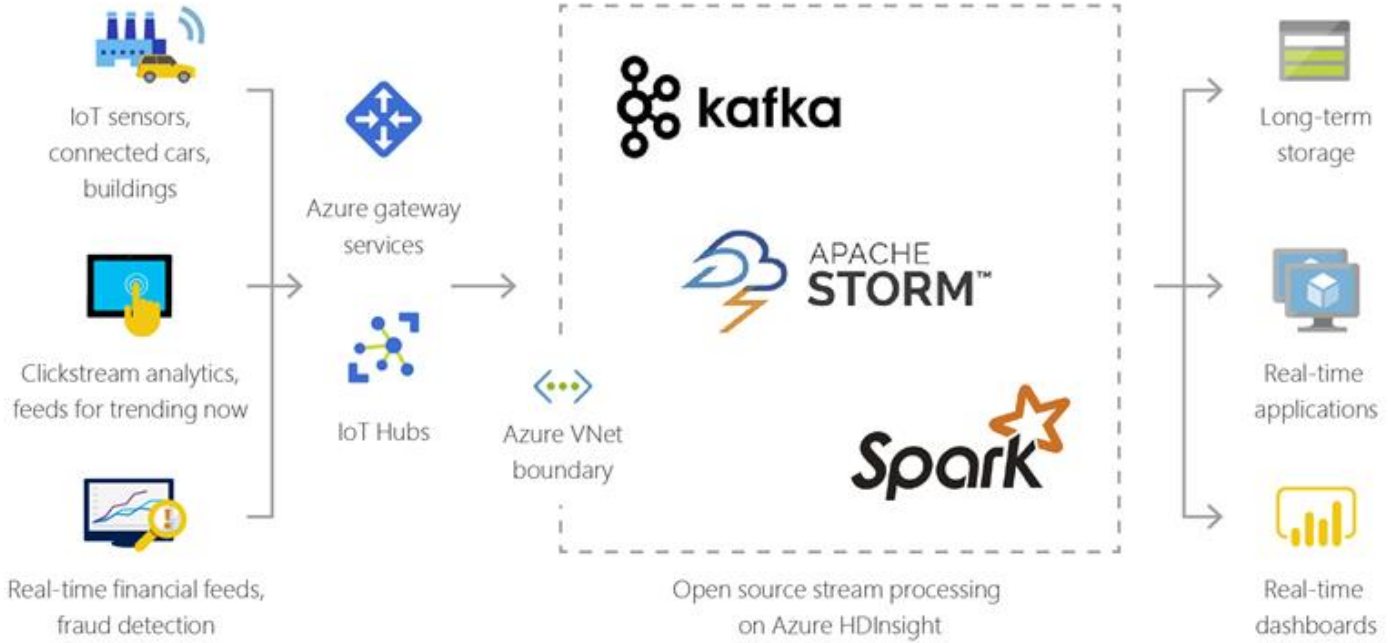
- **Transformacje** (ang. *transformations*) polegają na przekształceniach obiektów (czyli operacje Map), filtrowaniu listy obiektów, albo grupowaniu. Transformacje są wykonywane dopiero wtedy, kiedy pojawi się akcja, która wymaga dostępu do przetworzonych danych (leniwa ewaluacja).

- **Akcje** (*actions*) polegają np. na zliczaniu (czyli też z wszystkuim, co związane z operacją Reduce), zapisywaniu, albo wyświetlaniu przetworzonych danych.

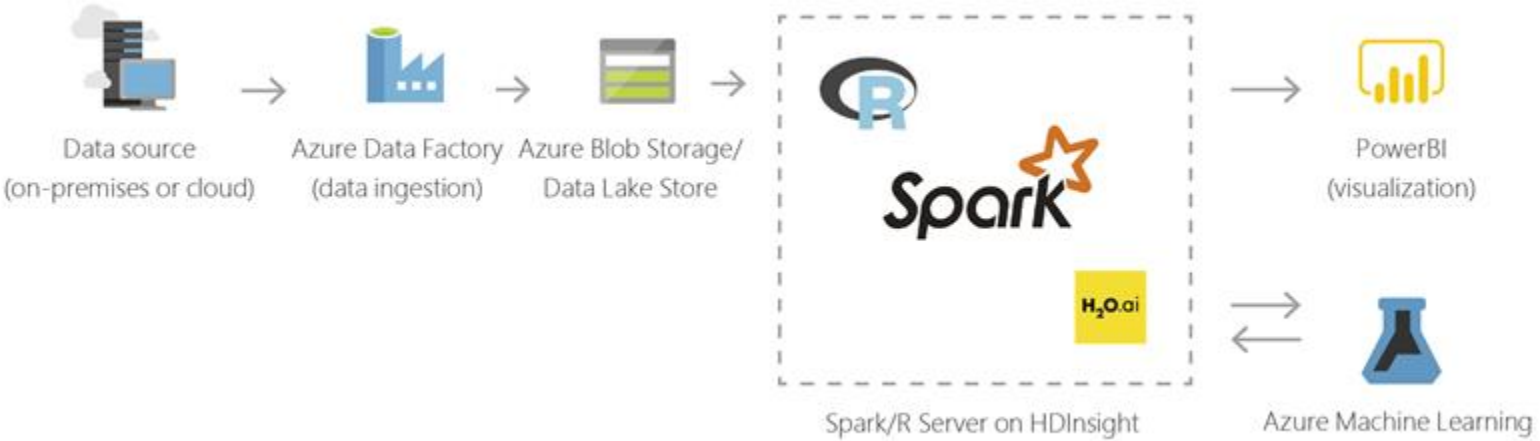
Spark - przypadki użycia

- Interakcyjna analiza danych i analiza biznesowa
- Spark Machine Learning
- Przesyłanie strumieniowe i analiza danych w czasie rzeczywistym na platformie Spark

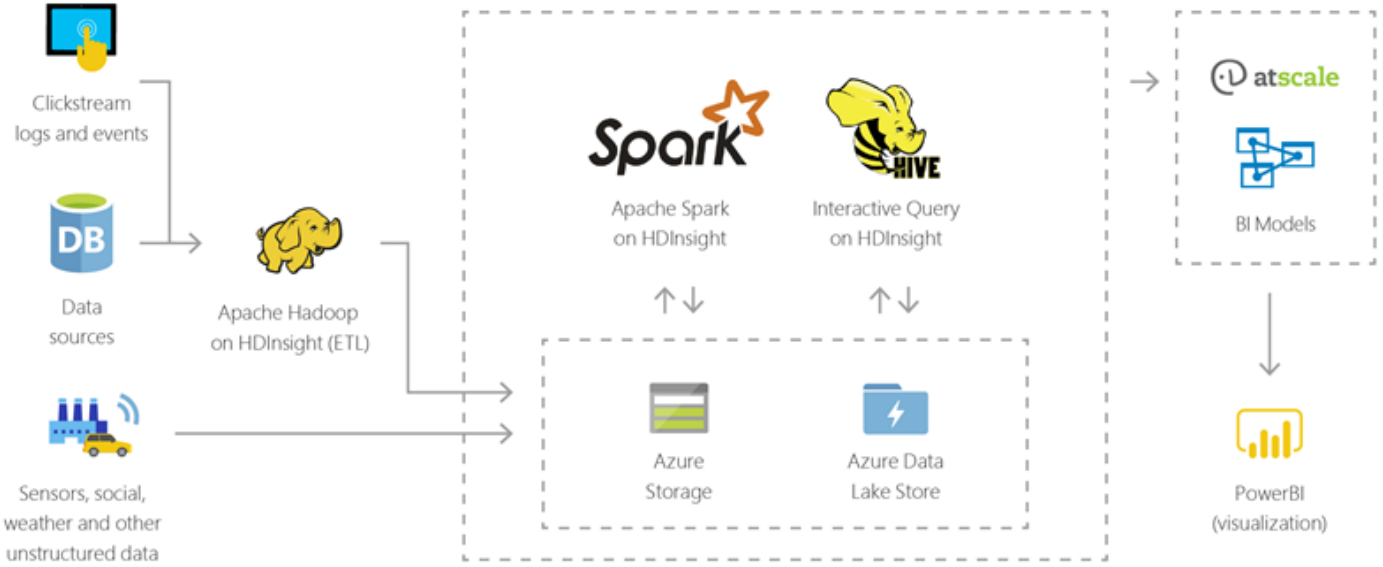
HDInsight - IoT



HDInsight – Data science



HDInsight – Data warehousing



HDInsight Spark – składniki klastra

- Spark Core, Spark SQL, GraphX, MLlib
- Anaconda
- Livy
- Notes Jupyter
- Notes Zeppelin

HDInsight Spark – magazyn danych

- **Azure Storage** – zalecany jako główny magazyn danych klastra (na nim zostanie utworzony klaster).
- **Azure Data Lake Storage** – zalecany jako dodatkowy storage na dane przetwarzane przez klaster ze względu na większą wydajność.

HDInsight Spark – format danych

- Text (csv, tsv, json…)
- Sequence File
- Avro
- Parquet
- ORC (Optimized Row Columnar)

Ćwiczenia

- Tworzenia klastra HDInsight Spark z poziomu witryny Azure Portal oraz uruchomienie interaktywnego notesu Jupyter.
- `
- Tworzenia klastra HDInsight Spark za pomocą skryptów oraz uruchomienie zadania za pośrednictwem interfejsu Livy.

Materiały

- https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-introduction
- https://docs.microsoft.com/en-us/azure/hdinsight
- https://spark.apache.org
- Advanced Analytics with Spark, 2nd Edition
- http://hadoop.apache.org
- https://hortonworks.com
- https://azure.microsoft.com/en-us/services/hdinsight
- https://azure.microsoft.com/en-us/services/hdinsight/apache-spark