Using Machine Learning with HDInsight

Eng Teong Cheah
MVP Visual Studio & Development Technologies

Agenda

Introduction to HDInsight

HDInsight and machine learning models

Introduction to HDInsight

Introducing Azure HDInsight

Azure HDInsight is a fully managed, full-spectrum, open-source analytics service for enterprises.

HDInsight is a cloud service that makes it easy, fast, and cost-effective to process massive amounts of data.

HDInsight also supports a broad range of scenarios, like extract, transform, and load (ETL); data warehousing; machine learning; and IoT.

HBase on HDInsight

Apache HBase is an open-source, NoSQL database that is built on Hadoop and modeled after Google BigTable. HBase provides random access and strong consistency for large amounts of unstructured and semistructured data in a schemaless database organized by column families.

HBase on HDInsight

Data is stored in the rows of table, and data within a row is grouped by column family.

HBase is a schemaless database in the sense that neither the columns nor the type of data stored in them need to be defined before using them.

The open-source code scales linearly to handle petabytes of data in thousands of nodes.

It can rely on data redundancy, batch processing, and other features that are provided by distributed applications in the Hadoop ecosystem.

Storm on HDInsight

Apache Storm is a distributed, fault-tolerant, open-source computation system.

You can use Storm to process streams of data in real time with Hadoop.

Storm solutions can also provide guaranteed processing of data, with the ability to replay data that was not successfully processed the first time.

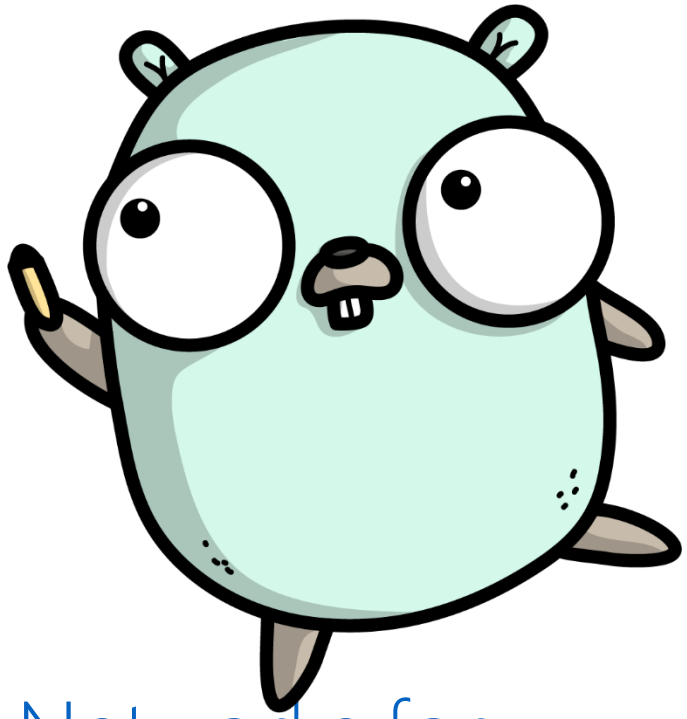
HDInsight and machine learning models

What is Spark

Spark provides primitives for in-memory cluster computing. A Spark job can load and cache data into memory and query it repeatedly.

In-memory computing is much faster than disk-based applications, such as Hadoop, which shares data through HDFS. Spark also integrates into the Scala programming language to let you manipulate distributed data sets like local collections. There's no need to structure everything as a map and reduce operations.

Resources



TutorialsPoint

Microsoft Docs

Lecture Collection | Convolutional Neural Networks for Visual Recognition(Spring 2017)

Python Numpy Tutorial

Image Credits: @ashleymcnamara

Thank you

Eng Teong Cheah
Microsoft MVP Visual Studio & Development Technologies
Twitter: @walkercet
Github: https://github.com/ceteongvanness
Blog: https://ceteongvanness.wordpress.com/
Youtube: http://bit.ly/etyoutubechannel