ROITRAINING

MAXIMIZE YOUR TRAINING INVESTMENT

Spark Program

CHAPTER 1: HDFS

Chapter Objectives

In this chapter, we will:

- Learn about the Hadoop Distributed Files System (HDFS)
- Run a standalone instance of HDFS
- Create directories and files in HDFS

About HDFS—I

- The Hadoop Distributed File System (HDFS) is the main storage used by Hadoop MapReduce applications
  - Distributed, POSIX-like file system
    - Designed to run on commodity hardware
    - Scales to clusters composed of thousands of nodes
  - Highly fault tolerant
    - Automatically detects hardware faults
    - Supports quick recovery
  - Implemented in Java
- Can be used as a standalone general purpose file system, but relaxes certain POSIX filesystem requirements
  - Designed for storing and reading very large files (>TB)
    - Supports high throughput read and writes
    - Write once, read many
    - Aimed at batch processing
    - Default block size is 128MB
  - Does not support random insertion or modification of data
  - Appending/truncating data is possible

About HDFS—II

- HDFS is used either directly or indirectly by many Big Data and NoSQL applications including:
  - Hadoop
  - Spark
  - HBase
  - Pig
  - Hive
  - Others

Core HDFS Services

- HDFS is implemented as several services which are usually deployed on a cluster of machines
  - Referred to as an HDFS cluster
  - Arranged in a controller/worker architecture
- Core HDFS services include:
  - **NameNode** (controller) stores file system metadata
  - **DataNode** (worker) stores file data (data blocks)
- The NameNode is the master server
  - Implements a POSIX-like hierarchical file system with '/' as the root directory
  - Enforces read/write permissions on files and directories
  - Tracks the location of the data blocks for each file
- The DataNode is the slave server
  - Handles read and write requests from HDFS clients
  - Performs block creation, deletion, and replication as instructed by the NameNode

Start Hadoop

- To start Hadoop on the VM
  - Open a terminal window and type the following commands:

```
sudu bash
start-hadoop
jps
```

- From a command line, enter the following commands:

```
hdfs
hdfs dfs
hdfs dfs -ls /
hdfs dfs -put ~/ROI/datasets/northwind/CSV/categories
/
hdfs dfs -ls /
```

Start Jupyter

- To start Jupyter with the latest lesson on the VM
  Open a terminal window and type the following commands:

```
cd ~/ROI
./start.sh
```

  - This will launch the browser so you can navigate to which lesson folder you want to work on

Command Line Examples

- As we have seen, HDFS provides a command line interface
- From a command line, enter the following commands:

```
hdfs
hdfs dfs
hdfs dfs -ls /
hdfs dfs -put ~/ROI/datasets/northwind/CSV/categories /
hdfs dfs -ls /
```

Chapter Summary

In this chapter, we have:

- Learned about the Hadoop Distributed Files System (HDFS)
- Ran a standalone instance of HDFS
- Created directories and files in HDFS