Hadoop

Introduction / Overview

Preface

- We will use these PowerPoint slides to guide us through our topic.

- Expect 15 minute segments of lecture

- Expect 1 - 4 hour lab segments

- Expect minimal pretty pictures

Objectives

- What is Hadoop?
  - Definition
  - Core Components
  - Software
    - Apache
    - Other
- Why do we need something like Hadoop?
- What skills do we need?
- Labs

What is Hadoop?

- Definition
  - Hadoop is an open source, Java-based programming framework that supports the processing and storage of extremely large data sets in a distributed computing environment. It is part of the Apache project sponsored by the Apache Software Foundation.

What is Hadoop?

- Core Components
  - *Hadoop Common* – contains libraries and utilities needed by other Hadoop modules
  - *Hadoop Distributed File System (HDFS)* – a distributed file-system that stores data on commodity machines, providing very high aggregate bandwidth across the cluster
  - *Hadoop YARN* – a platform responsible for managing computing resources in clusters and using them for scheduling users' applications
  - *Hadoop MapReduce* – an implementation of the MapReduce programming model for large-scale data processing

What is Hadoop?

- Software (Apache):

  - **Pig** - A platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs.

  - **Hive** - A data warehouse software project built on top of Hadoop for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop. Hive is now being deprecated.

  - **Beeline** – A wrapper around Hive. JDBC based and more secure than Hive

  - **HBase -** A column-oriented key/value data store built to run on top of the Hadoop Distributed File System (HDFS).

  - **Spark -** Spark is a fast and general processing engine compatible with Hadoop data designed to perform both batch processing (similar to MapReduce) and new workloads like streaming, interactive queries, and machine learning.

What is Hadoop?

- Software (Apache):
  - **Zeppelin -** An open web-based notebook that enables interactive data analytics.
  - **ZooKeeper -** A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.
  - **Flume -** A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
  - **Sqoop -** A tool designed for efficiently transferring bulk data between Hadoop and structured datastores such as relational databases.
  - **Oozie -** A server-based workflow scheduling system to manage Hadoop jobs.

What is Hadoop?

- Software (Apache):
  - **Storm -** A free and open source distributed real-time computation system. Storm makes it easy to reliably process unbounded streams of data, doing for real-time processing what Hadoop did for batch processing.
  - **HCatalog** - A metadata abstraction layer that insulates users and scripts from how and where data is physically stored. Used primarily by Pig, MapReduce, and Hive.
    - HCatLoader – Interface for reading from HCatalog table
    - HCatStorer – Interface for writing to HCatalog table
  - **WebHCat** - A component that provides a set of REST-like APIs for HCatalog and related Hadoop components.

What is Hadoop?

- Software Other (Hadoop Distributions)
  - **Cloudera** (Open Source with some proprietary components)
    - Cloudera Manager (Management Interface)
    - Impala (SQL Interface)
    - Cloudera Search (Product search and access)
  - **Hortonworks** (Open Source)
    - Ambari (Management Interface - Apache)
    - Stinger (Query Interface)
    - Apache Solr (data searching)
  - **MapR** (Proprietary File System, MapRFS)

What is Hadoop?

- Software Other (Cloud Services)
  - Microsoft Azure
  - Amazon AWS
  - Many others

Why do we need something like Hadoop?

- The Hadoop framework provides tools for efficiently accessing mammoth sets of data. Hadoop is used to push code to data which is fragmented across clusters of disk drives.

- The framework reduces data processing time at a percentage based on the number of drives the data is clustered across.

- The framework supports built-in data redundancy and protection from hardware failure.

What skills do we need?

- **Jav**a – The Hadoop framework is based on Java. If you’re not familiar with Java you can still use other tools to access Hadoop data.

- **Python** – A great language for mapping and / or reducing data. A great language for both Hadoop and Spark development.

- **Scala** – The native language of Spark.

- **SQL** – Hive and Beeline like you to know this (as do other Hadoop technologies)

What skills do we need?

- **UNIX / Linux**
  - **vi, vim, etc.** - for editing scripts.
  - **Scripting** – For wrapping and launching code written in various languages.
  - **Aliases** – For giving more user friendly names to your Hadoop commands.
  - **Data streams** – For understanding the flow of data from application to
  - application.
  - **Pipes** – For capturing and filtering stream data.
  - **Redirection** – For storing stream data.
  - **awk** – For robust scripting capabilities.

Labs

- Set up and Practice (Lab 1)
  - Ambari Overview
  - Putty Setup
  - Accessing the AWS UNIX box
  - Checking software
    - Java
    - Python
  - Hadoop – A few commands
  - UNIX Overview
    - Processes
    - environment
    - .bash_profile
    - .bashrc
    - alias setup