BIG DATA LAB ASSIGNMENT TWO

Devansh Sanghvi
Department Of Biotechnology
Indian Institute of Technology, Madras
Chennai, India
be19b002@smail.iitm.ac.in

***Abstract*—This document aims to answer the two questions provided in Assignment Two of the course: Big Data Lab**

I. Question Two

Question: Provide a brief description of the functionality of the following services:

- HDFS
- Hive
- Pig
- Yarn

Solution

HDFS

1) HDFS is a distributed file system that stores and manages large amounts of data across multiple machines in a cluster. It is an important part of the Hadoop ecosystem and is intended to be highly scalable and fault-tolerant.
2) HDFS is built on a master-slave architecture, with a single NameNode acting as the master and managing the file metadata, and multiple DataNodes acting as slaves and storing the actual data.
3) HDFS is optimised for large file handling and is intended to provide high throughput rather than low-latency data access.
4) This is accomplished by partitioning files into blocks and distributing these blocks across multiple DataNodes. Each block is duplicated several times to ensure fault tolerance and availability.

Hive

1) Hive is a data warehouse system that provides a SQL-like interface known as HiveQL for querying and analysing Hadoop data.
2) HiveQL is similar to SQL, but it is designed for large-scale datasets and complex data processing tasks. The Hive compiler converts HiveQL into MapReduce jobs, which are then executed on the Hadoop cluster.
3) Hive supports a wide range of data sources and file formats, including structured data, semi-structured data, and unstructured data. It also includes the Hive Metastore, a metadata repository that stores information about tables, columns, and partitions, making it easier to manage large datasets.

Pig

1) Pig Latin is a high-level language that provides a platform for analysing large datasets. Pig Latin is a scripting language designed to abstract the underlying complexity of MapReduce programming and provide a more intuitive and concise means of expressing data analysis programmes.
2) The Pig compiler converts Pig Latin programmes into MapReduce jobs, which are then executed on the Hadoop cluster.
3) Pig includes a wide range of built-in data functions and operators, as well as the ability to define custom functions in Java or Python. Pig also provides a data flow model that allows users to specify the dependencies between different processing steps, making it easier to write complex data analysis pipelines.

Yarn

1) Yarn is a framework for managing and scheduling Hadoop cluster resources. It provides a centralised platform for managing resources such as CPU, memory, and disc, allowing different applications to coexist and efficiently share the same resources.
2) Yarn enables Hadoop to support a variety of data processing applications such as MapReduce, Spark, and Tez.
3) Yarn is built on a master-slave architecture, with a single Resource Manager acting as the master and managing resource allocation to various applications and multiple Node Managers acting as slaves and managing resources on individual machines.
4) Yarn has a pluggable architecture that allows different scheduling algorithms to be used depending on the application's needs. This enables the allocation of resources for various purposes to be optimised.