**Multi-node Hadoop Cluster Using Cloudera**

CSE 487

Cloud Computing

Summer 172

Department of Computer Science and Engineering

United International University

Submitted by: Group 4

Group members

1. Asif Ahmed - [aahmed141068@bscse.uiu.ac.bd](mailto:aahmed141068@bscse.uiu.ac.bd)
2. Md. Younus Bipul - [mbipul141075@bscse.uiu.ac.bd](mailto:mbipul141075@bscse.uiu.ac.bd)
3. Syed Md. Imran - [simran141086@bscse.uiu.ac.bd](mailto:simran141086@bscse.uiu.ac.bd)
4. Niger Sultana Tahniat - [ntahniat141088@bscse.uiu.ac.bd](mailto:ntahniat141088@bscse.uiu.ac.bd)
5. Toha Khan Mozlish - [tmozlish141089@bscse.uiu.ac.bd](mailto:tmozlish141089@bscse.uiu.ac.bd)

Multi-node Cluster

Multi node or Fully Distributed Cluster is a typical hadoop cluster which follows a master-slave architecture. It will basically comprise of one master machine (running the NameNode and TaskTracker daemon) and one or more slave machines (running the DataNode and TaskTracker daemon). The default replication factor for a multi node cluster is 3. It is basically used for full stack development of hadoop application and projects.

Hadoop

Apache Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the [MapReduce](https://en.wikipedia.org/wiki/MapReduce) programming model. It consists of computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures are common occurrences and should be automatically handled by the framework.

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to. This allows the dataset to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are distributed via high-speed networking

Hive

Apache Hive is a data warehouse software project built on top of Apache 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. Traditional SQL queries must be implemented in the [MapReduce](https://en.wikipedia.org/wiki/MapReduce) Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries ([HiveQL](https://en.wikipedia.org/wiki/HiveQL" \o "HiveQL)) into the underlying Java without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop.[]](https://en.wikipedia.org/wiki/Apache_Hive#cite_note-:3-3) While initially developed by Facebook, Apache Hive is used and developed by other companies such as Netflix and the Financial Industry Regulatory Authority (FINRA). Amazon maintains a software fork of Apache Hive included in Amazon Elastic MapReduce on Amazon Web Services

VirtualBox

A VirtualBox or VB is a software virtualization package that installs on an operating system as an application. VirtualBox allows additional operating systems to be installed on it, as a Guest OS, and run in a virtual environment. In 2010, VirtualBox was the most popular virtualization software application. Supported operating systems include Windows XP, Windows Vista, Windows 7, [macOS X](https://www.computerhope.com/jargon/m/macosx.htm), [Linux](https://www.computerhope.com/jargon/l/linux.htm), Solaris, and OpenSolaris. The current version is 5.1.28 . It is a type-2 hypervisor that sits on an host OS and can run multiple gues OS in it.

Cloudera Manager

Coudera manager is a software that makes it easy to manage Hadoop deployments of any scale in production. Quickly deploy, configure and monitor your cluster through an intuitive UI-complete with roling upgrades, backups and disaster recovery and customizable alerting. Cloudera manager is available as integrated and supported part of Cloudera Enterprise. The current version is Cloudera Manager 5.12.1 .