Tuesday, May 3, 2022

1. Installing Hadoop (see git hup for instructions)

To install hadoop make sure java is already installed and then follow git instructions

“jps” - to list the Hadoop stricture

Linux, text editor (nano, gedit- good but doesn’t work on remote, echo command, chmod command

1. Installing HIVE (see git hub for instructions)

Data warehouses build on top of hadoop,

Meta data engine + query engine for Hadoop

1. Big data: collection of large datasets that include huge volumes, high velocity, and variety of data that is increased day by day
2. Using traditional data management systems, it’s difficult to process big data

Why:

Only vertical scalability

1. Hadoop contains two modules:

**MapReduce:** it’s parallel programing model for processing large amounts of structured, Semi Structured, and unstructured data

**HDFS:**

1. Hadoop ecosystem contains different sub-projects(tools) such as

Sqoop:

Pig:

**Hive:** data warehouse infrastructure tool to process structured data in Hadoop

It resides on top of Hadoop to summarize Big Data, and make querying and analyzing

* Hive is not OLTP,

<https://github.com/nodesense/cts-data-engineering-feb-2022/tree/main/notes>