Task1

Explain the below concepts with an example in brief.

● Nosql Databases

Term NoSQL stands for Not Only SQL. NoSQL is a database design that can store variety of data models such as document, key-value, graphical and columnar format. It is an alternative to traditional DB's where data is stored in tables and data schema is designed before a DB is created. NoSQL databases are very much useful for working with large sets of distributed data.

Advantages of NoSQL

- More Scalable
- Superior performance
- Simplicity of design
- It can handle huge volumes of structured, semi-structured, and unstructured data
- Object-oriented programming which is easy to use and flexible
- Scaling to cluster of machines in its architecture instead of expensive, monolithic architecture

Examples: HBase, MongoDB, Cassandra

● Types of Nosql Databases

Key-Value model:- Data is represented as a collection of key-value pairs and is the simplest data model in nosql.

Eg: Oracle NoSql, Dynamo

Document data model:- Value of each key is a document. This model can contain key – values, key-arrays and nested documents.

Eg: Mongo Db, IBM Domino

Graph model:- Data relations are represented in graph model with elements interconnected with some relations between them. Eg: road maps, network topology data

Eg: Neo4J

Columnar model:- Data is stored as rows of transactions, for fast retrieval.

Eg: AWS, HBase

● CAP Theorem

CAP theorem states that a distributed computing system will not be able to provide consistency, availability and partition tolerance at the same time.

C Stands for **consistency** which means when performing an activity same information must be received irrespective of the node that processed it, In other words, all clients see the same data at the same time.

A stands for **Availability** which means system is available to all clients for read and write operation.

P Stands for **Partition tolerance** which means that the system will function even if there is a message loss or failure in among any partitioned servers.

● HBase Architecture

Hbase has master-slave architecture with three servers such as HMaster, HRegionServer and Zookeeper.

HMaster

HMaster acts as the master server in the master slave architecture. It coordinates the Hbase clustersand is responsible for operations of the cluster. If any of the region servers connected to it fails, HMaster assigns another available region server. HMaster can also assign a region to another region server as part of load balancing.

HRegions Servers

HRegionServer hosts and manages regions and splits regions automatically. It handles all read and write requests coming from clients and communicates with clients directly. For each column family, HRegions maintain a store. Main component of HRegions is Memstore - which holds in-memory modifications to the store Hfile

Zookeeper

Zookeeper is a centralized monitoring server that maintains configuration information and provides distributed synchronization. If the client wants to communicate with regions servers, client has to approach Zookeeper.

● HBase vs RDBMS

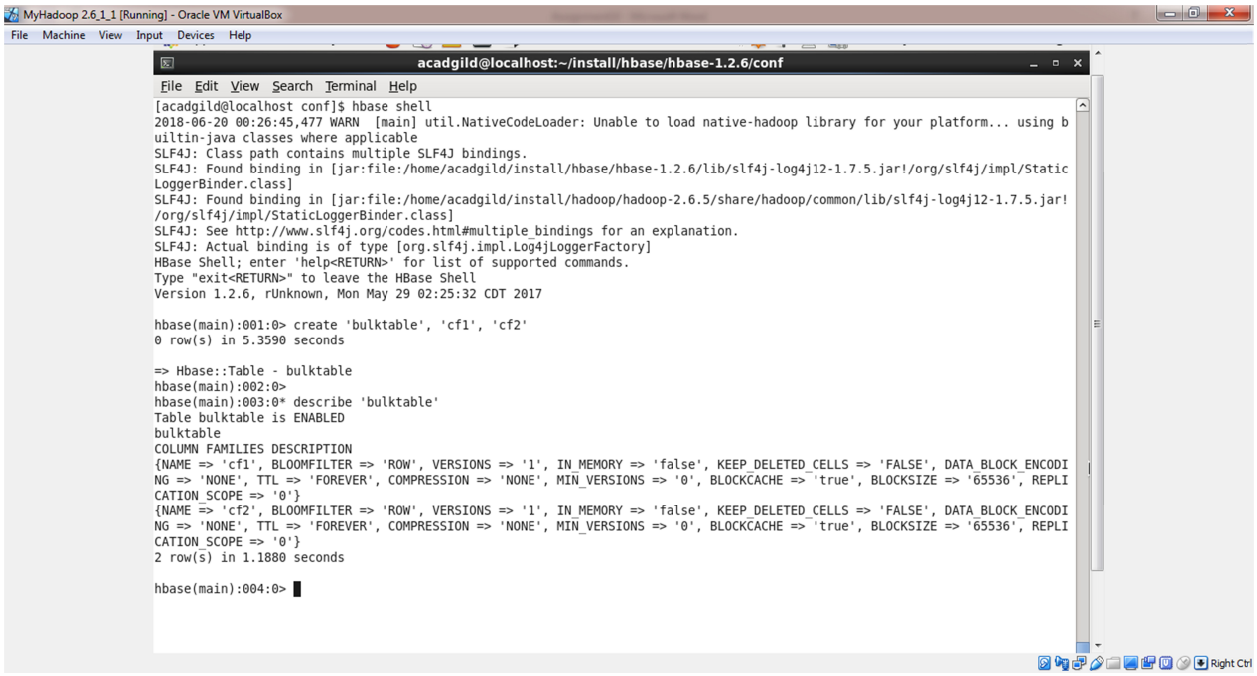
HBASE	RDBMS
Column oriented, suitable for OLAP	Row oriented, suitable for OLTP
Can store structured, semi-structured or de-normalized data	Stores only Structured, normalized data
Enables aggregation over many rows and Columns	Aggregation is an expensive operation in RDBMS
Can store large volumes of data	Cannot store large volumes like HBASE
Open source a[ache product	Licensing cost is very high

Task2

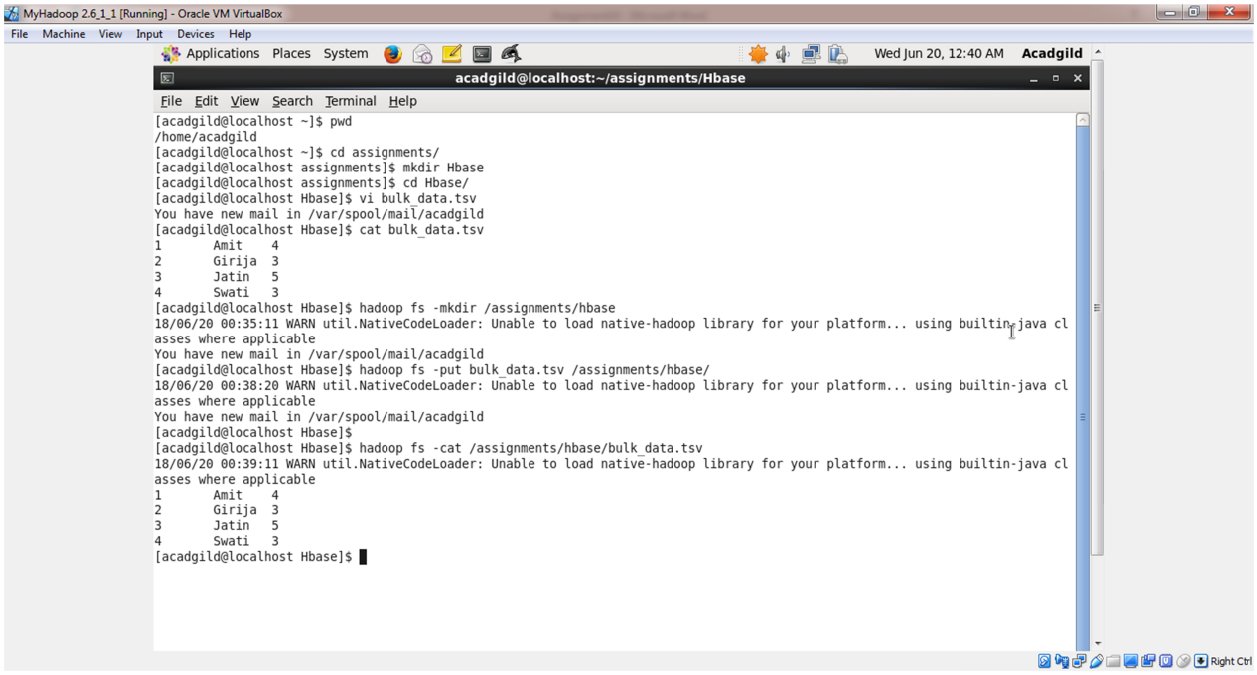
Execute blog present in below link

https://acadgild.com/blog/importtsv-data-from-hdfs-into-hbase/

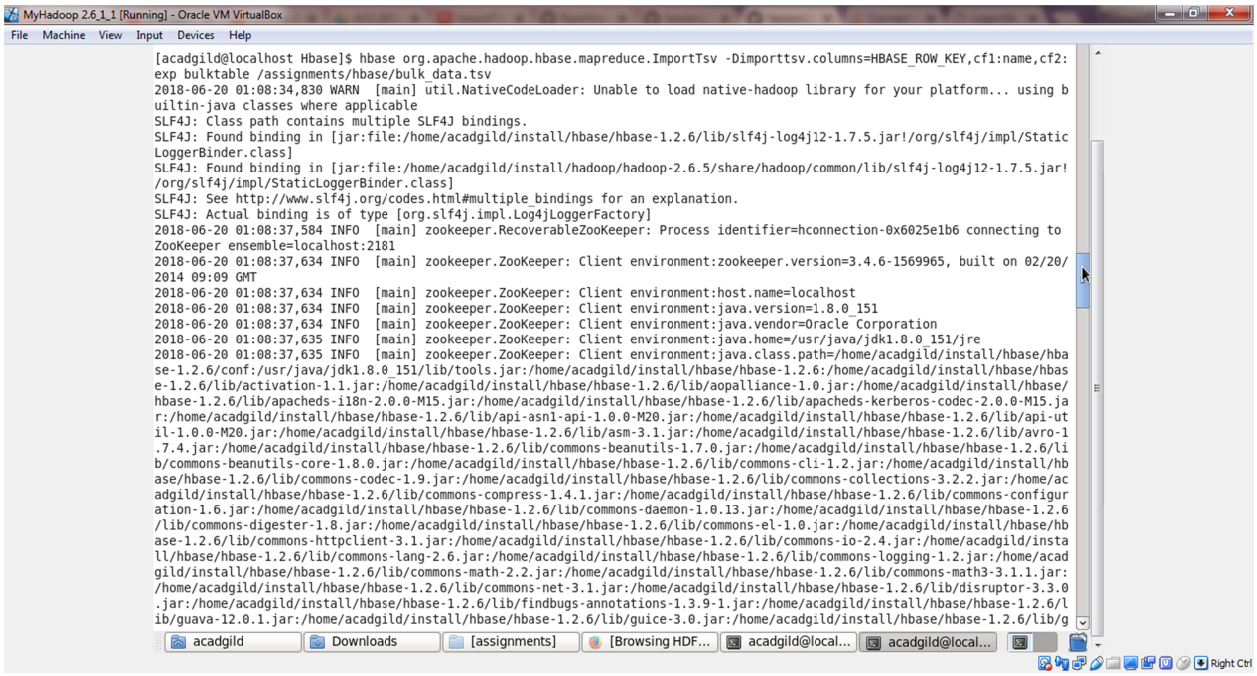
Create a table bulk_table with two column families cf1 and cf2.

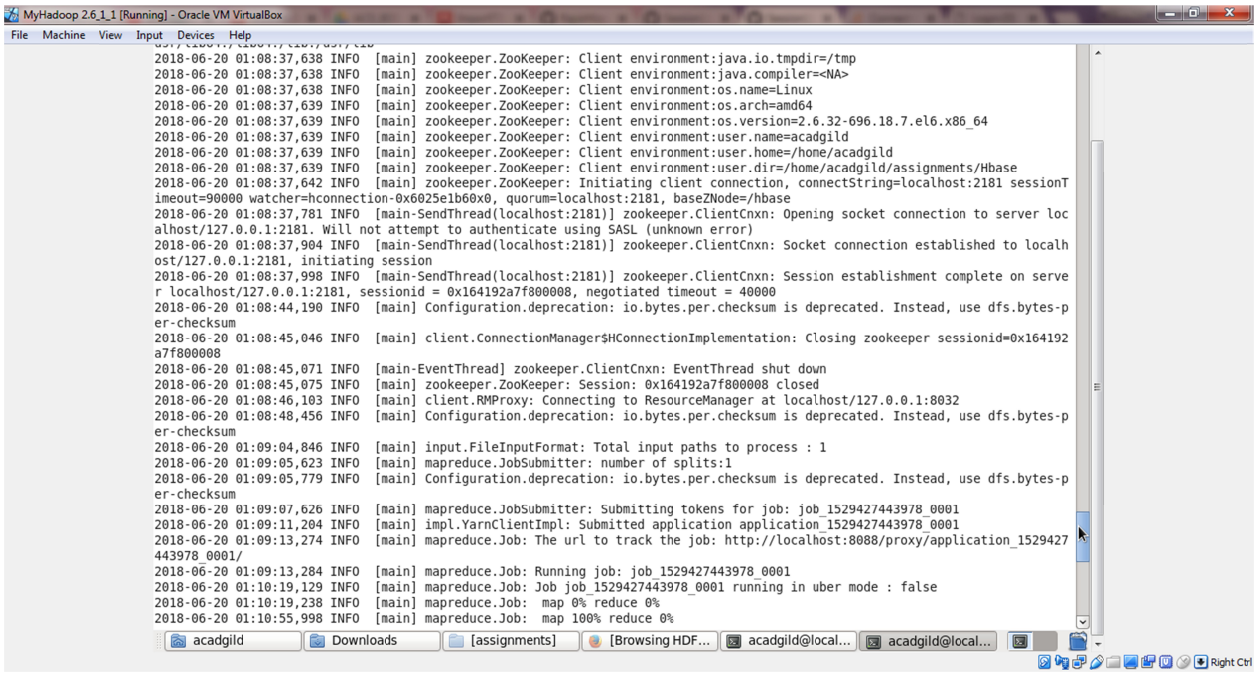


Create a file bulk_table.tsv with data delimited by tabs and load the file to HDFS.



$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.columns=a,b,c <tablename> <hdfs-inputdir> is the command which takes data from HDFS and loads into Hbase.


MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
[acadgild@localhost Hbase]$ hbase org.apache.hadoop.hbase.mapreduce.ImportTsv -Dimporttsv.columns=HBASE_ROW_KEY,cf1:name,cf2:exp bulktable /assignments/hbase/bulk_data.tsv
2018-06-20 01:08:34,830 WARN [main] util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/home/acadgild/install/hbase/hbase-1.2.6/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/home/acadgild/install/hadoop/hadoop-2.6.5/share/hadoop/common/lib/slf4j-log4j12-1.7.5.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
2018-06-20 01:08:37,584 INFO [main] zookeeper.RecoverableZooKeeper: Process identifier=hconnection-0x6025e1b6 connecting to ZooKeeper ensemble=localhost:2181
2018-06-20 01:08:37,634 INFO [main] zookeeper.ZooKeeper: Client environment:zookeeper.version=3.4.6-1569965, built on 02/20/2014 09:09 GMT
2018-06-20 01:08:37,634 INFO [main] zookeeper.ZooKeeper: Client environment:host.name=localhost
2018-06-20 01:08:37,634 INFO [main] zookeeper.ZooKeeper: Client environment:java.version=1.8.0_151
2018-06-20 01:08:37,634 INFO [main] zookeeper.ZooKeeper: Client environment:java.vendor=Oracle Corporation
2018-06-20 01:08:37,635 INFO [main] zookeeper.ZooKeeper: Client environment:java.home=/usr/java/jdk1.8.0_151/jre
2018-06-20 01:08:37,635 INFO [main] zookeeper.ZooKeeper: Client environment:java.class.path=/home/acadgild/install/hbase/hbase-1.2.6/conf:/usr/java/jdk1.8.0_151/lib/tools.jar:/home/acadgild/install/hbase/hbase-1.2.6:/home/acadgild/install/hbase/hbase-1.2.6/lib/activation-1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/aopalliance-1.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-i18n-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/apacheds-kerberos-codec-2.0.0-M15.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-asn1-api-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/api-util-1.0.0-M20.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/asm-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/avro-1.7.4.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-beanutils-1.7.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-beanutils-core-1.8.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-cli-1.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-codec-1.9.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/comnons-collections-3.2.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-compress-1.4.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-configuration-1.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-daemon-1.0.13.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-digester-1.8.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-el-1.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-httpclient-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-io-2.4.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-lang-2.6.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-logging-1.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math-2.2.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-math3-3.1.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/commons-net-3.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/disruptor-3.3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/findbugs-annotations-1.3.9-1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guava-12.0.1.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/guice-3.0.jar:/home/acadgild/install/hbase/hbase-1.2.6/lib/g
acadgild
Downloads
[assignments]
[Browsing HDF...
acadgild@local...
acadgild@local...
Right Ctrl


MyHadoop 2.6.1_1 [Running] - Oracle VM VirtualBox
File Machine View Input Devices Help
2018-06-20 01:08:37,638 INFO [main] zookeeper.ZooKeeper: Client environment:java.io.tmpdir=/tmp
2018-06-20 01:08:37,638 INFO [main] zookeeper.ZooKeeper: Client environment:java.compiler=<NA>
2018-06-20 01:08:37,638 INFO [main] zookeeper.ZooKeeper: Client environment:os.name=Linux
2018-06-20 01:08:37,639 INFO [main] zookeeper.ZooKeeper: Client environment:os.arch=amd64
2018-06-20 01:08:37,639 INFO [main] zookeeper.ZooKeeper: Client environment:os.version=2.6.32-696.18.7.el6.x86_64
2018-06-20 01:08:37,639 INFO [main] zookeeper.ZooKeeper: Client environment:user.name=acadgild
2018-06-20 01:08:37,639 INFO [main] zookeeper.ZooKeeper: Client environment:user.home=/home/acadgild
2018-06-20 01:08:37,639 INFO [main] zookeeper.ZooKeeper: Client environment:user.dir=/home/acadgild/assignments/Hbase
2018-06-20 01:08:37,642 INFO [main] zookeeper.ZooKeeper: Initiating client connection, connectString=localhost:2181 sessionTimeout=90000 watcher=hconnection-0x6025e1b60x0, quorum=localhost:2181, baseZNode=/hbase
2018-06-20 01:08:37,781 INFO [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Opening socket connection to server localhost/127.0.0.1:2181. Will not attempt to authenticate using SASL (unknown error)
2018-06-20 01:08:37,904 INFO [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Socket connection established to localhost/127.0.0.1:2181, initiating session
2018-06-20 01:08:37,998 INFO [main-SendThread(localhost:2181)] zookeeper.ClientCnxn: Session establishment complete on server localhost/127.0.0.1:2181, sessionid = 0x164192a7f800008, negotiated timeout = 40000
2018-06-20 01:08:44,190 INFO [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-06-20 01:08:45,046 INFO [main] client.ConnectionManager$HConnectionImplementation: Closing zookeeper sessionid=0x164192a7f800008
2018-06-20 01:08:45,071 INFO [main-EventThread] zookeeper.ClientCnxn: EventThread shut down
2018-06-20 01:08:45,075 INFO [main] zookeeper.ZooKeeper: Session: 0x164192a7f800008 closed
2018-06-20 01:08:46,103 INFO [main] client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
2018-06-20 01:08:48,456 INFO [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-06-20 01:09:04,846 INFO [main] input.FileInputFormat: Total input paths to process : 1
2018-06-20 01:09:05,623 INFO [main] mapreduce.JobSubmitter: number of splits:1
2018-06-20 01:09:05,779 INFO [main] Configuration.deprecation: io.bytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2018-06-20 01:09:07,626 INFO [main] mapreduce.JobSubmitter: Submitting tokens for job: job_1529427443978_0001
2018-06-20 01:09:11,204 INFO [main] impl.YarnClientImpl: Submitted application application_1529427443978_0001
2018-06-20 01:09:13,274 INFO [main] mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1529427443978_0001/
2018-06-20 01:09:13,284 INFO [main] mapreduce.Job: Running job: job_1529427443978_0001
2018-06-20 01:10:19,129 INFO [main] mapreduce.Job: Job job_1529427443978_0001 running in uber mode : false
2018-06-20 01:10:19,238 INFO [main] mapreduce.Job: map 0% reduce 0%
2018-06-20 01:10:55,998 INFO [main] mapreduce.Job: map 100% reduce 0%
acadgild
Downloads
[assignments]
[Browsing HDF...
acadgild@local...
acadgild@local...
Right Ctrl

Check if the data is loaded correctly to HBase.