**TY B-TECH (CSE) – II [ 2023-24]**

**6CS371: Advanced Database System Lab.**

**Assignment No. 10**

**Cassandra Clustering**

**Name: Trilok Kulkarni**

**PRN: 21510119 TY-CSE**

* **Title:** Comprehensive Guide to Setting Up a Multi-Node Cassandra Cluster and Designing a Database Schema for IoT Temperature Sensor Data
* **Objective:** The aim of this guide is to provide comprehensive instructions for setting up a multi-node Cassandra cluster, installing and configuring DataStax OpsCenter, designing an efficient database schema tailored for storing weather station IoT temperature sensor data, and demonstrating cluster operations utilizing OpsCenter.
* **Theory and Concepts:** Clustering, in the context of Cassandra, is the fundamental mechanism for distributing and organizing data across multiple nodes within a cluster. Cassandra, being a distributed database system, prioritizes high availability and scalability, with clustering playing a pivotal role in achieving these objectives.
* **Data Distribution:** In a Cassandra cluster, data distribution is managed through partitioning, where each node is tasked with storing a specific portion of the dataset. Through intelligent partitioning strategies, Cassandra ensures even distribution of data across the cluster, optimizing performance and fault tolerance.
* **Replication:** To bolster fault tolerance and ensure data availability, Cassandra employs replication techniques, wherein copies of data are stored across multiple nodes. The replication factor, a configurable parameter, dictates the number of copies maintained and their placement across the cluster, enhancing resilience against node failures.
* **Scalability:** Cassandra's architecture facilitates horizontal scalability, allowing the cluster to expand seamlessly by adding more nodes. This scalability feature empowers organizations to accommodate growing data volumes and increasing workloads without encountering downtime or performance bottlenecks.
* **Comprehensive Guide Overview:**
* **Setting Up a Multi-Node Cassandra Cluster:**

Installation of Cassandra on multiple nodes.

Configuration of cluster settings, including seeds and communication protocols.

Validation of cluster connectivity and node synchronization.

* **Installing and Configuring DataStax OpsCenter:**

Deployment of OpsCenter for centralized cluster management and monitoring.

Configuration of monitoring thresholds and alerting policies.

Integration with existing authentication mechanisms for enhanced security.

Designing a Database Schema for Weather Station IoT Temperature Sensor Data:

* **Analysis of data requirements and access patterns.**

Development of an optimized schema leveraging Cassandra's data modeling principles.

Consideration of data partitioning and replication strategies to ensure efficient data storage and retrieval.

* **Demonstrating Cluster Operations Using OpsCenter:**

Utilizing OpsCenter's intuitive interface for cluster administration tasks.

Performing routine maintenance operations, such as node additions or removals.

Monitoring cluster health metrics and diagnosing performance issues.

**Multi node cluster on single machine**

**Steps :**

1. Install the Cassandra instance on the machine
2. Create a folder for storing the 3 instances of the Cassandra
3. Create 3 folders ( 3 nodes ) and copy the Cassandra instances in each of the folder
4. For each of the node in the cluster repeat the following process

* Go to config folder of the that node and open Cassandra.yaml file
* Edit the listen address , rpc address as to **127.0.0.1 , 127.0.0.2 and 127.0.0.3** on each of the node
* Edit the seeds with “**127.0.0.1 , 127.0.0.2, 127.0.0.3** ”
* Configure the ports on each of the node

1. Then test the cluster using **cassandra.bat -f** in 1st node
2. Then start the instances using **Cassandra -f** in each node
3. To see the status of the cluster run command **: nodetool status**
4. Now cluster is ready to use

**Setting up cluster on multiple machine**

* Here the most of the process remains the same just we need to put the actual ip addresses of the corresponding nodes in cluster in place of rpc address and listen address for each of node in cluster
* And edit seed with list of all the ip addresses of the nodes in cluster which will be basically a comma separated list
* Before setting up the Cassandra cluster we must have python 2.7.17 version , jdk8 and Apache Cassandra 3.11.4 version

1. **Configuration details of each node**

* **Node 1 :**

port : 7199

rpc\_address : 127.0.0.1

listen\_address : 127.0.0.1

seeds : 127.0.0.1, 127.0.0.2, 127.0.0.3

* **Node 2 :**

port : 7299

rpc\_address : 127.0.0.2

listen\_address : 127.0.0.2

seeds : 127.0.0.1, 127.0.0.2, 127.0.0.3

* **Node 3 :**

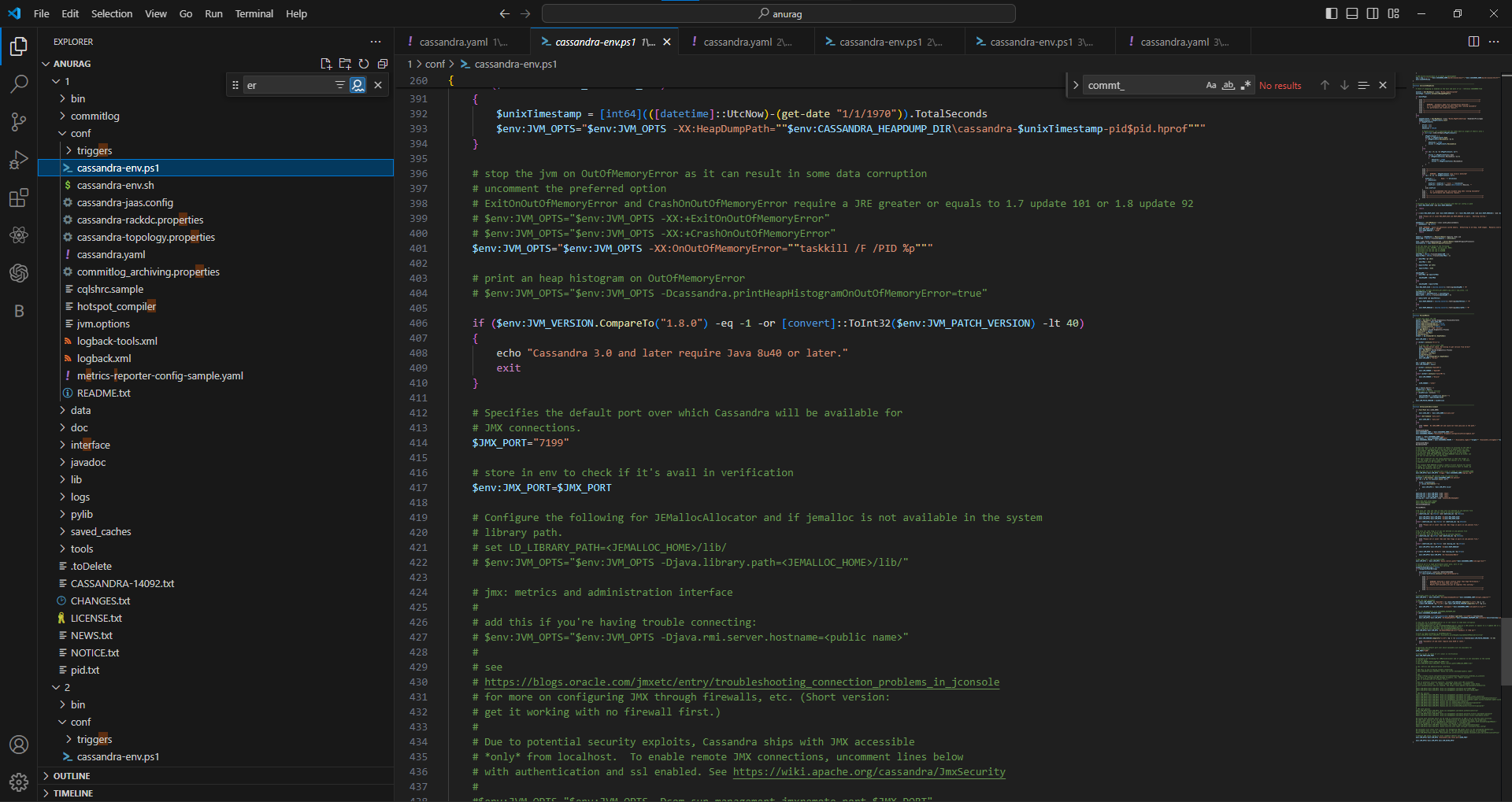
port : 7399

rpc\_address : 127.0.0.3

listen\_address : 127.0.0.3

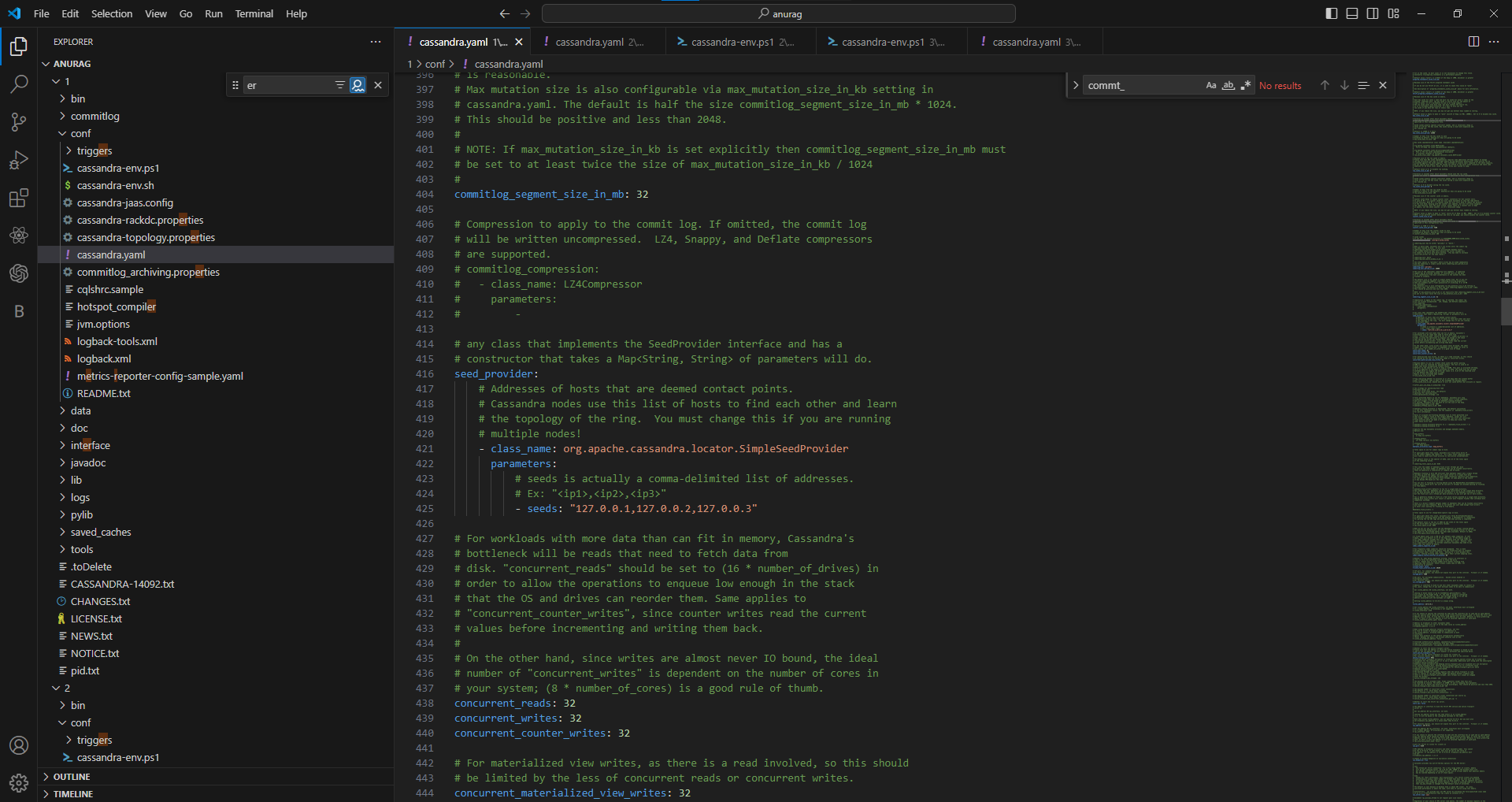
seeds : 127.0.0.1, 127.0.0.2, 127.0.0.3

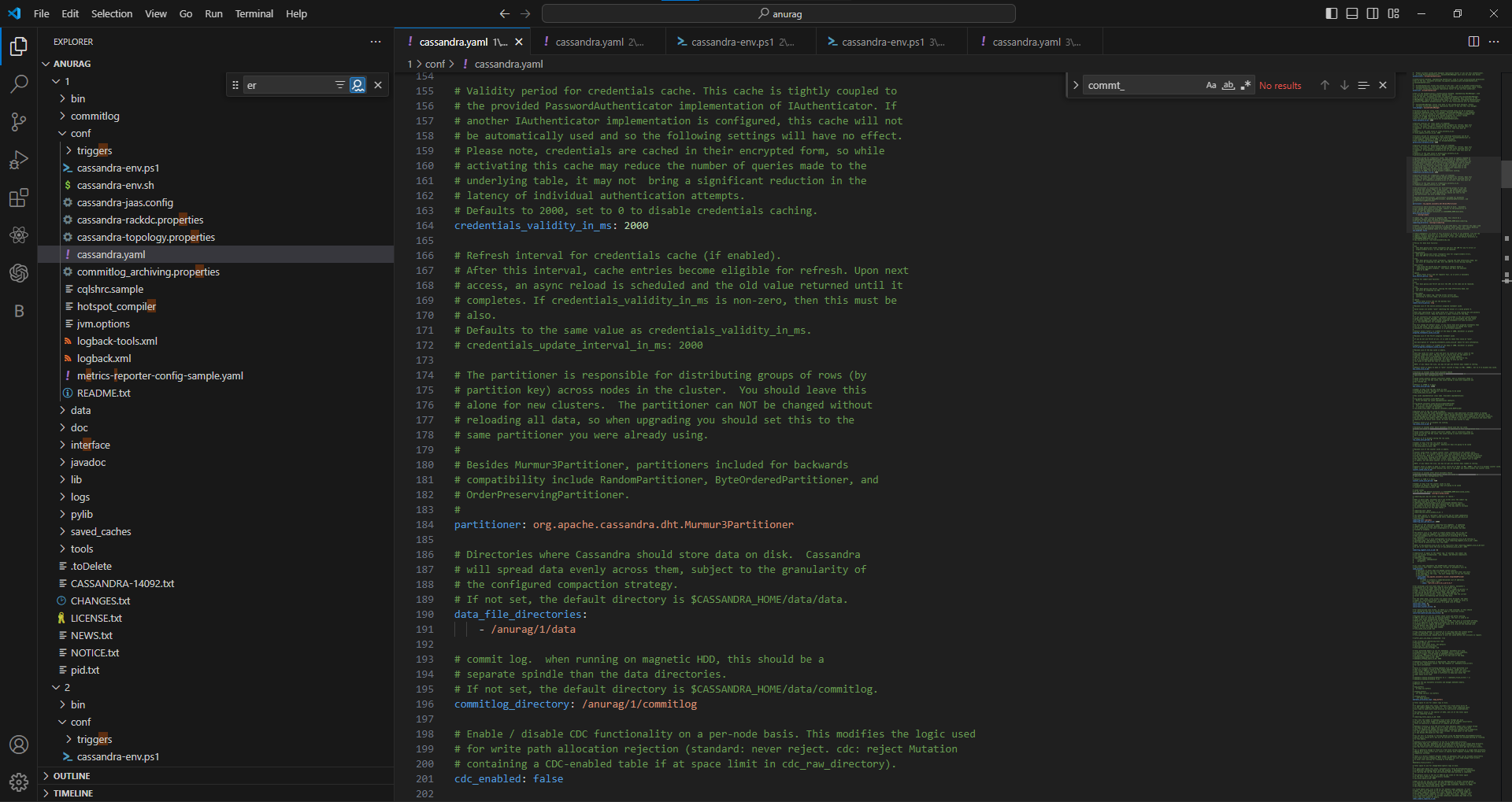
**Setting up cluster name**

****

Specifying the RPC\_Address , Setting up listen address and Setting up the seeds for all three nodes then Setting up the ports For node 1 : 7199, For node 2: 7299, For node 3: 7399 And similarly we need to edit ports in cassandra.bat file for each node Now we need to update the commit log directory , data file directory , saved cache directory of each of the node according to the path of individual node

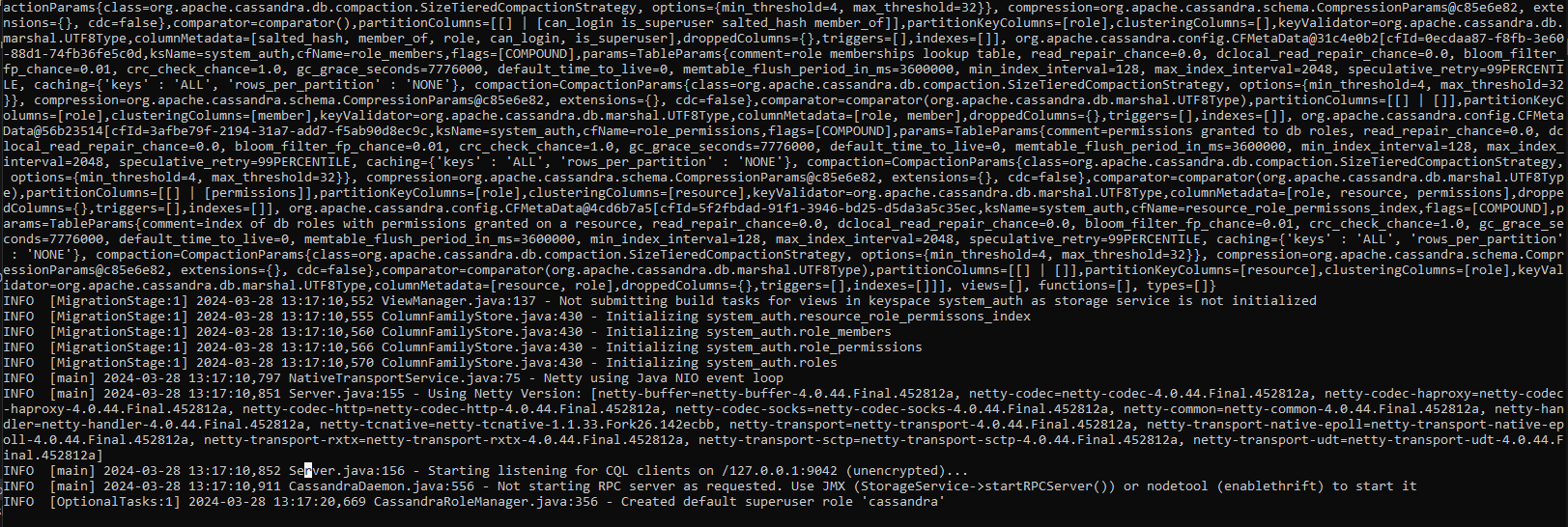
From Cassandra.yaml file

****

****

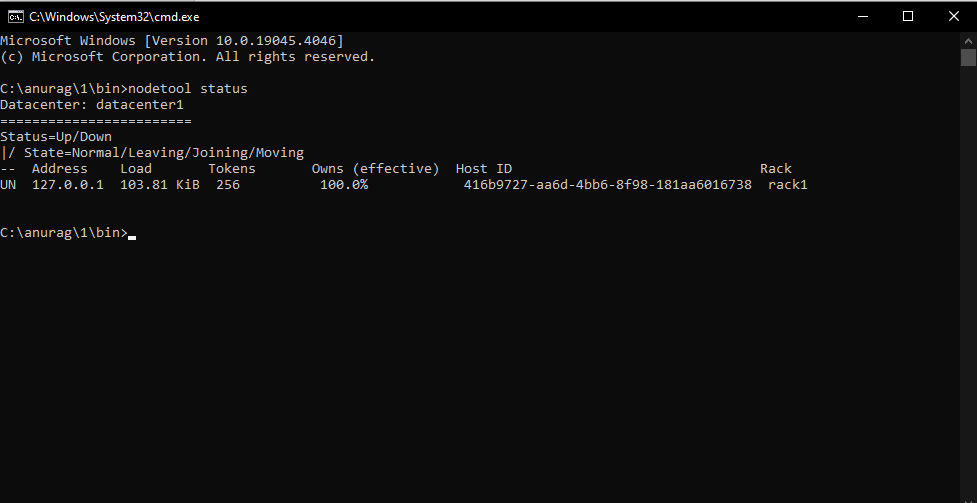
**Starting the node 1**

**Command : Cassandra -f**

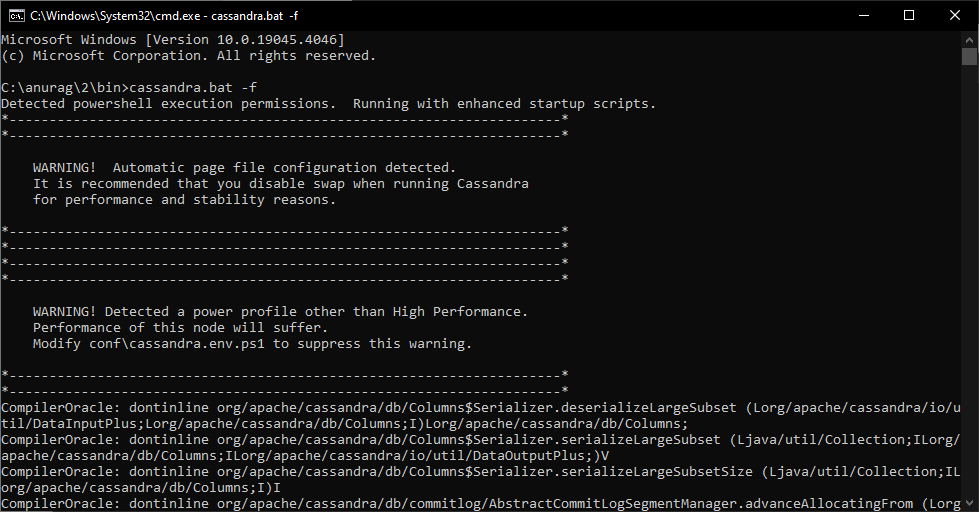
****

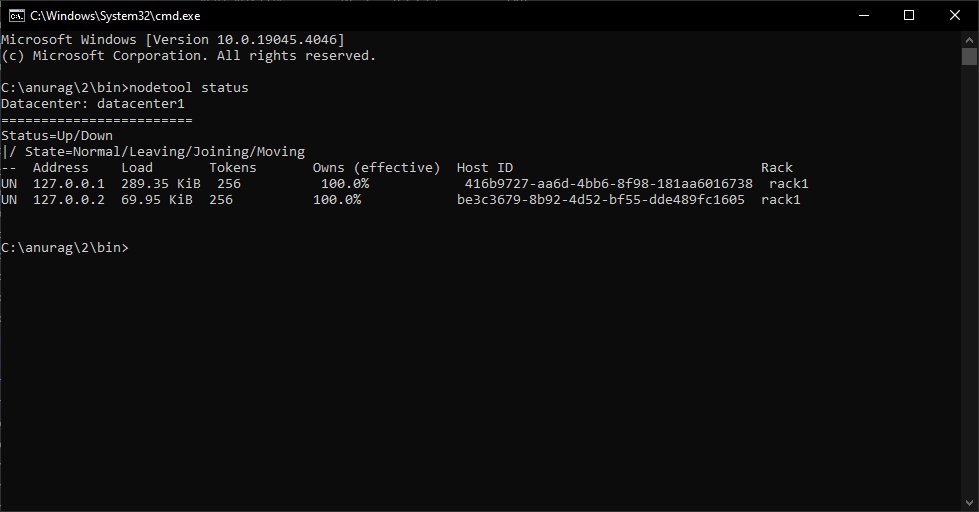
**Checking status after node 1 started :**

**Command : nodetool status**

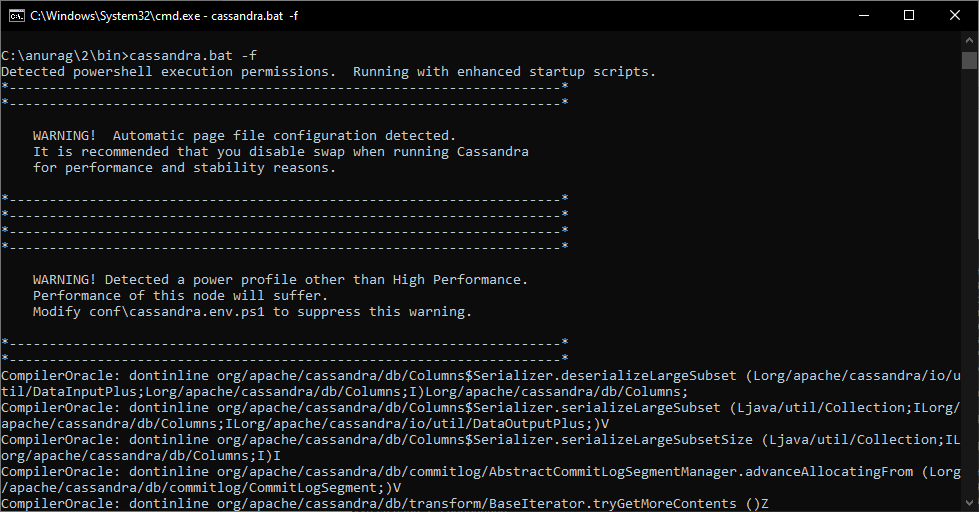
****

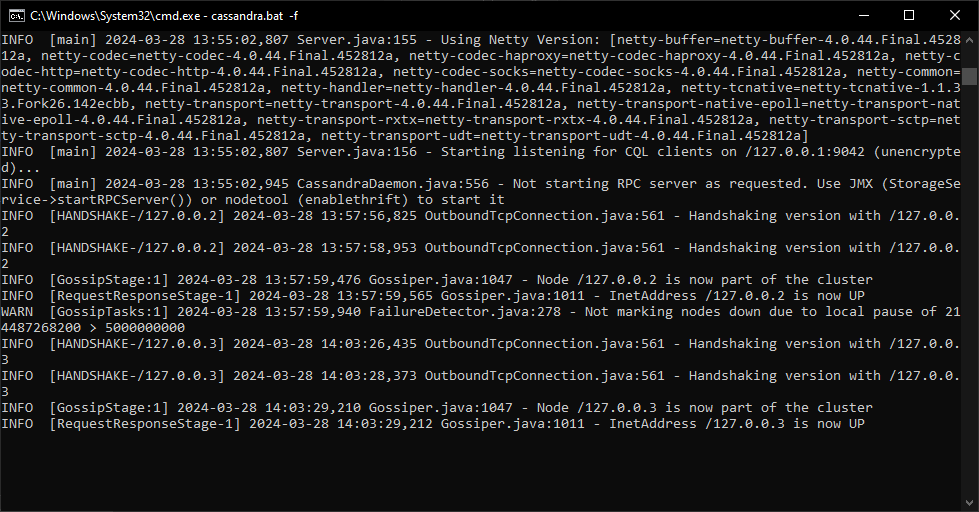
**Node 2 started**

****

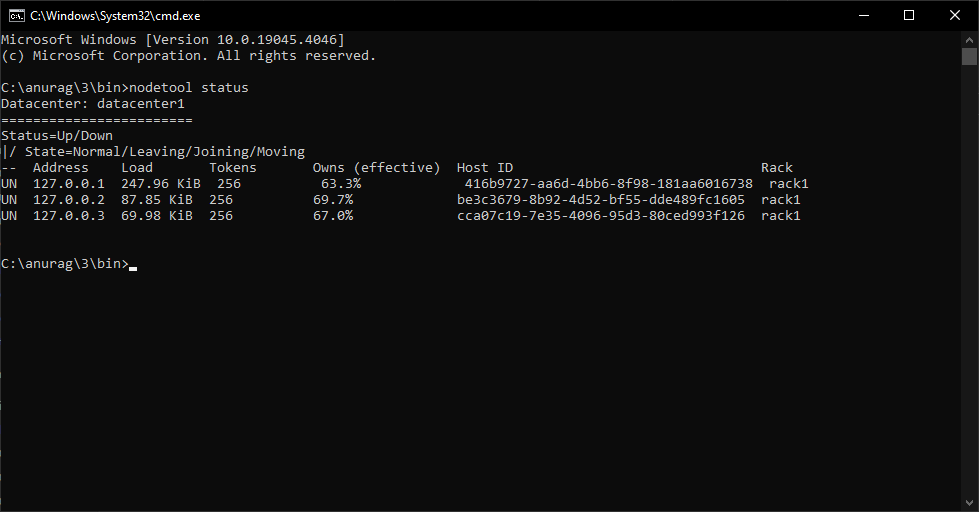
****

**Node 3 started**

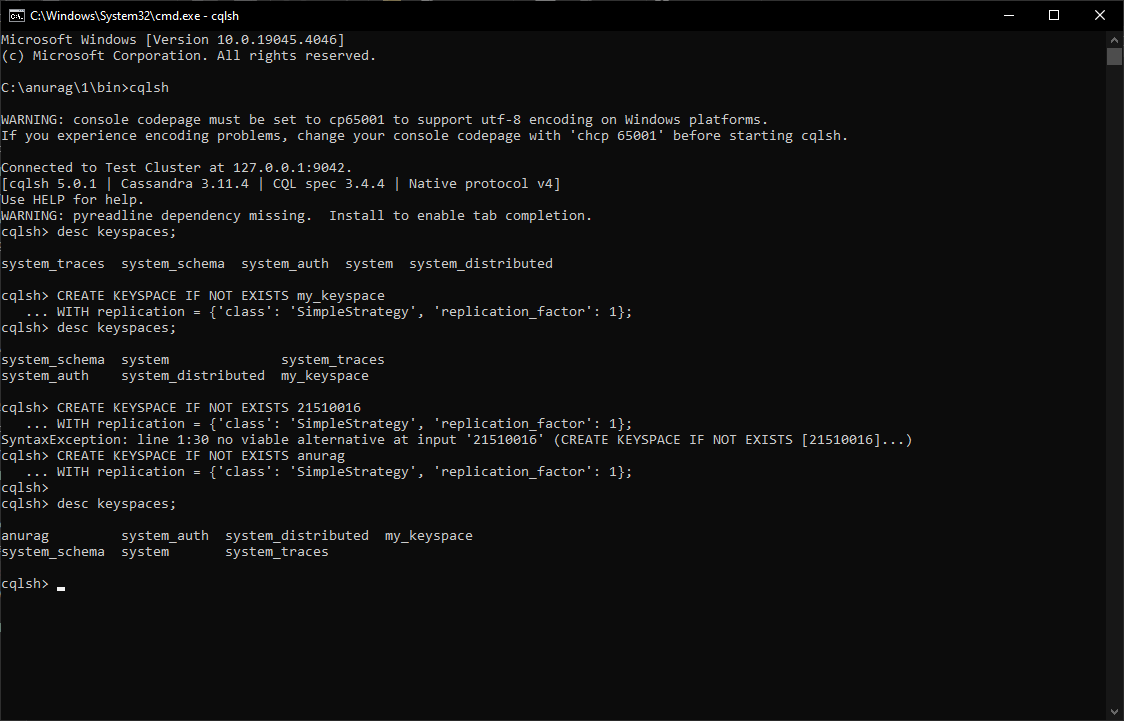
****

****

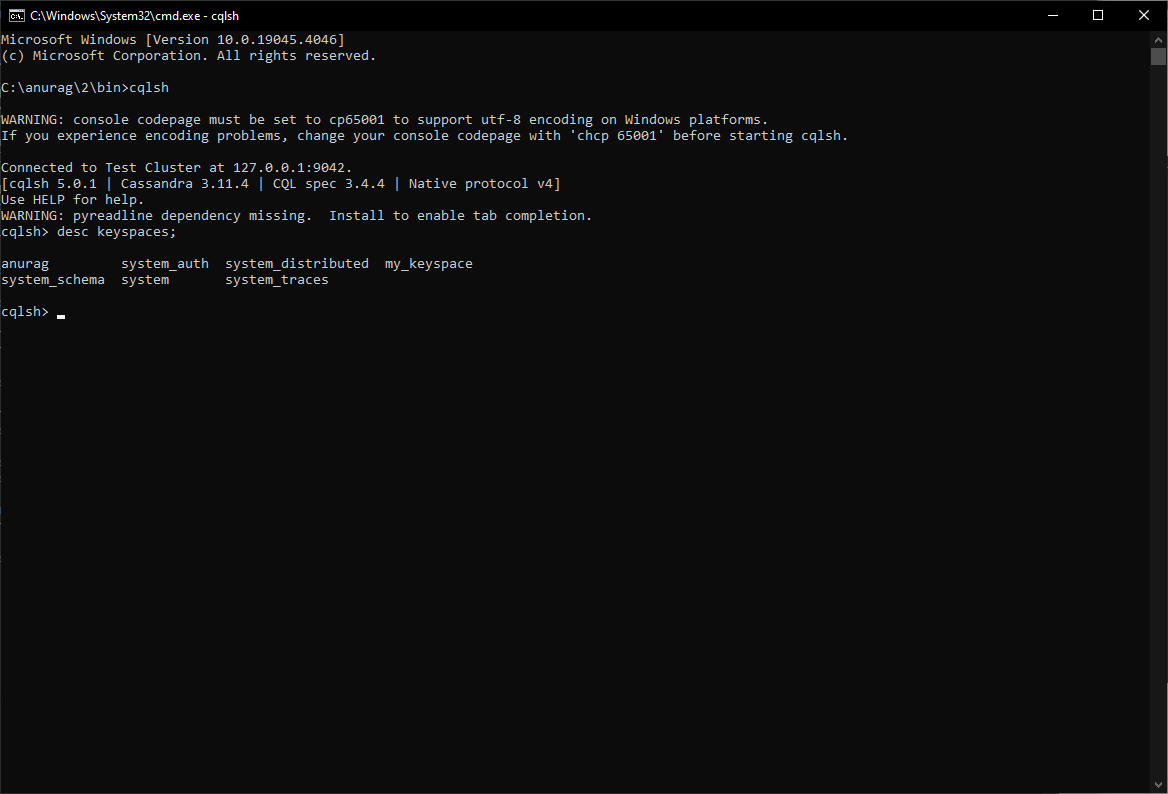
**Status after starting all nodes**

****

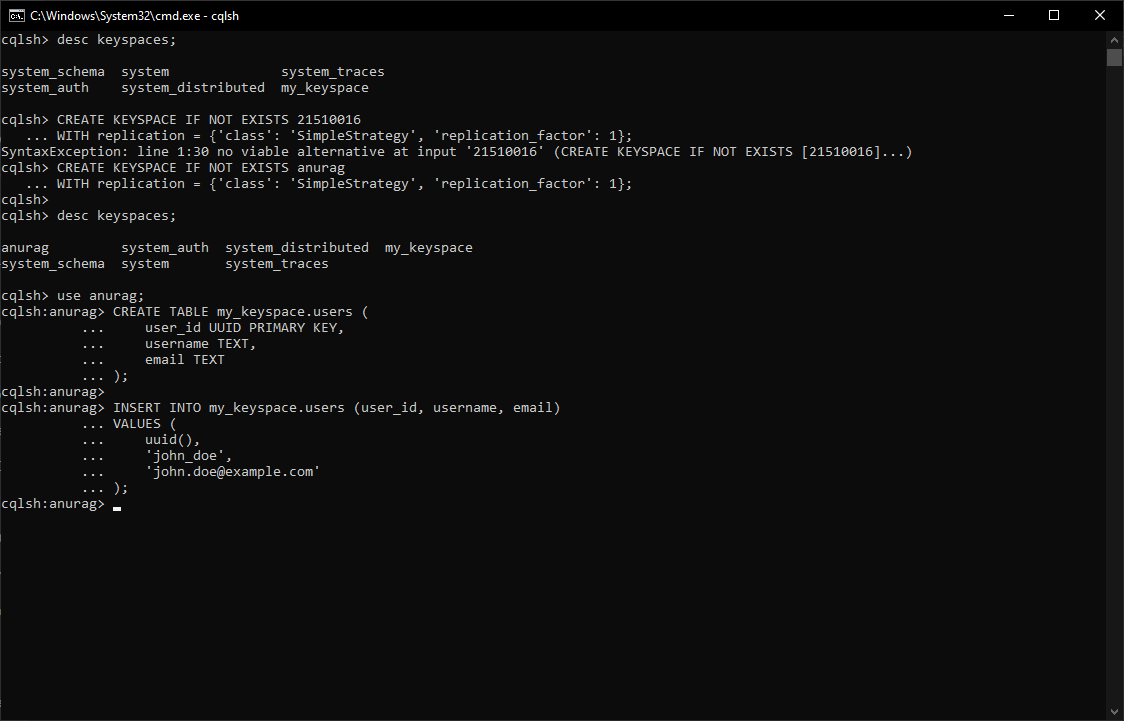
**Creating keyspace on node 1**

****

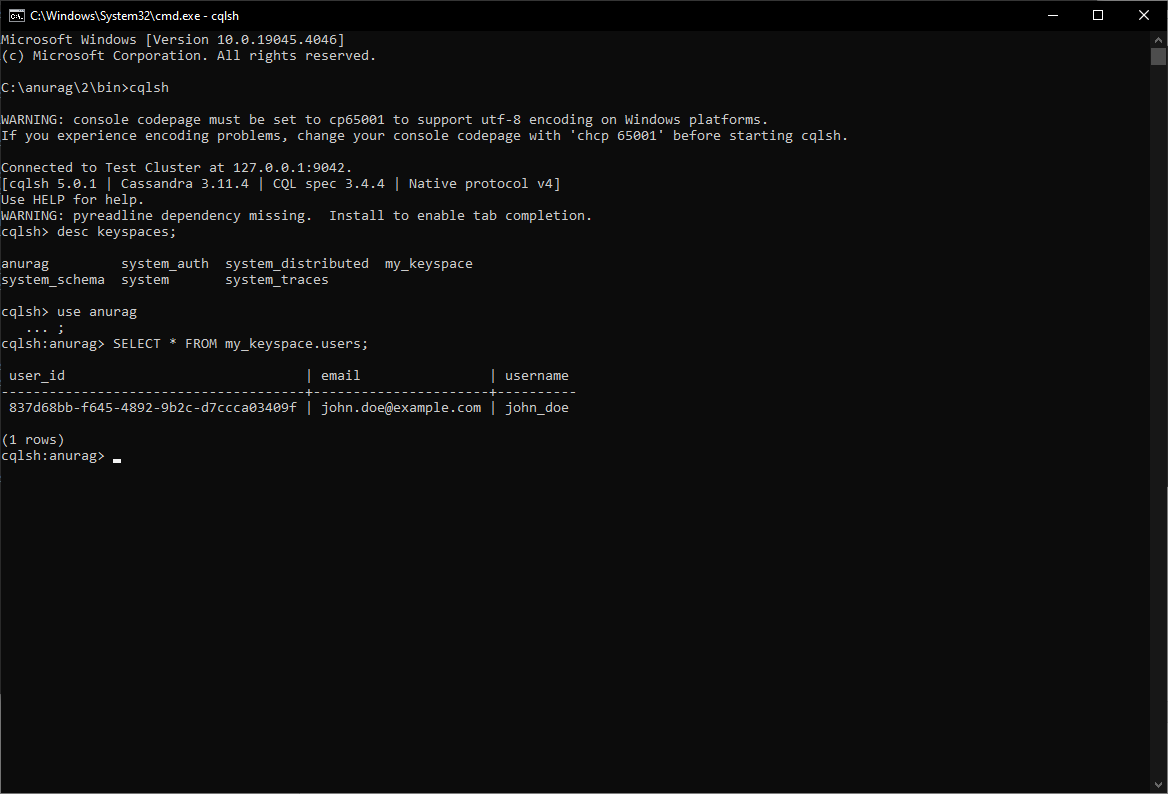
**Keyspace visible at other nodes**

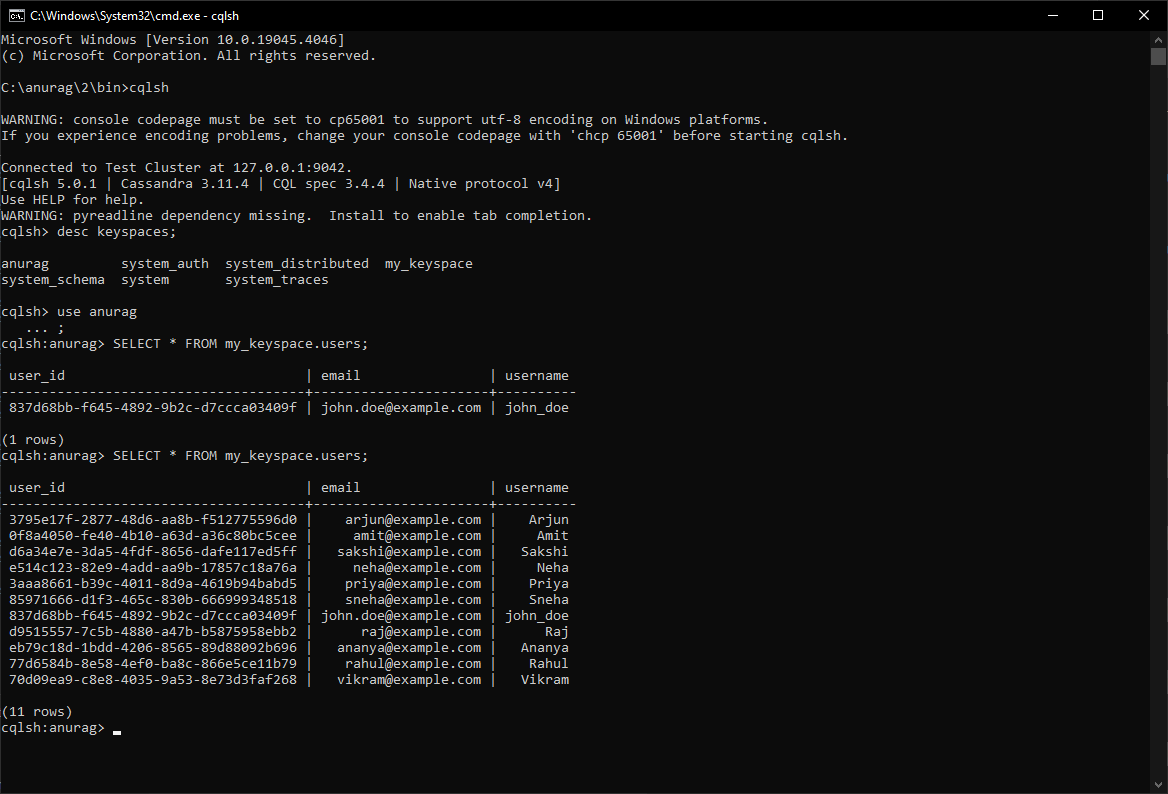
****

**Table created on node 1**

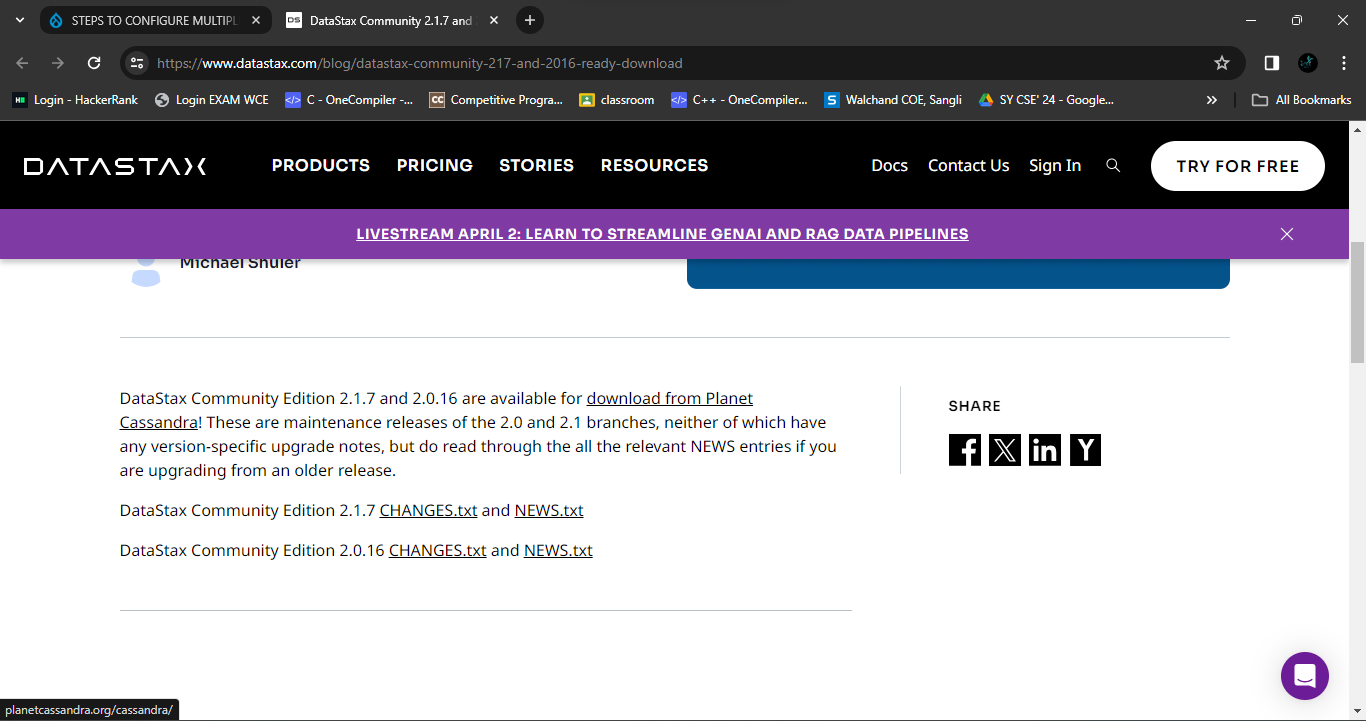
****

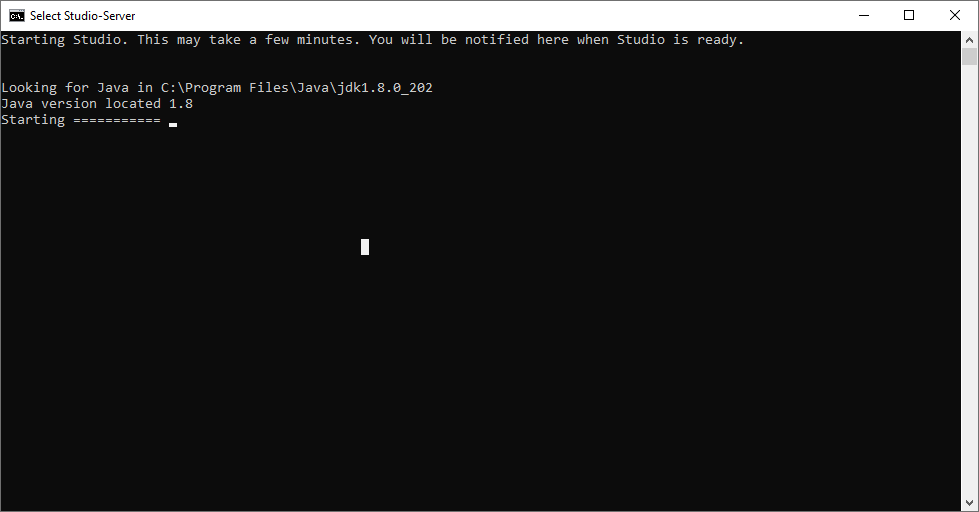
**Tables visible at other nodes**

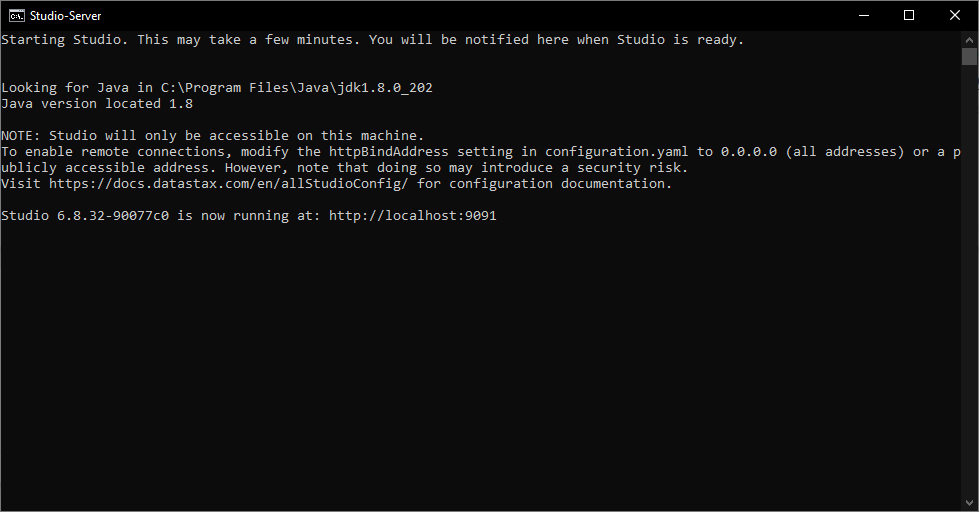
****

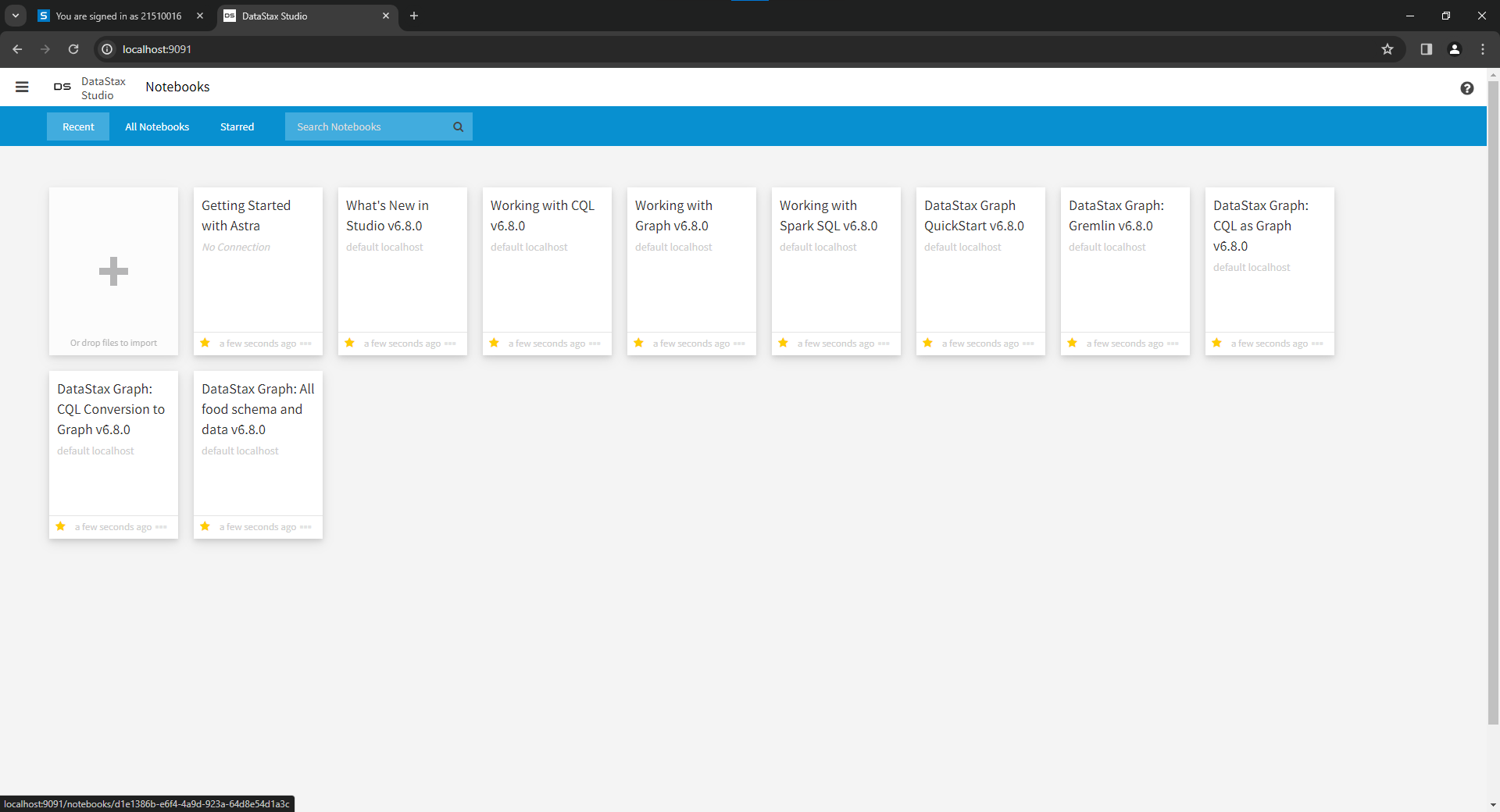
****

**Datastax installation**





****

****

**Conclusion :**   
In summary, this assignment encapsulates the comprehensive process of establishing a multi-node Cassandra cluster, deploying and fine-tuning DataStax OpsCenter, while delving into essential concepts such as Cassandra clustering and pivotal network configuration parameters including rpc\_address, listen\_address, and seeds. By engaging in practical tasks and hands-on exercises, participants not only grasp the fundamentals of distributed database systems but also delve into intricate facets of data modelling, proactive monitoring, and proficient cluster management. This holistic approach fosters a deeper understanding of the intricacies involved in orchestrating and optimizing Cassandra clusters for real-world applications, equipping learners with invaluable skills essential for managing robust distributed database infrastructures.