

Advanced Data Management - Assignment N°02

The assignment sheet contains one exercise with seven tasks. Last chance of uploading the assignments will be 16.11.2023. For each task, there is a maximum allotted time. Please do not work longer than the specified time on a task! If you think that you will not be able to finish the task in the given maximum time, stop working on it 10 minutes before the end, and provide an explanation containing the following information:

- Whether you think that the task is solvable with the current system at all, and why?
- If you think that is solvable with more time, which approach would you try out next?

Note: Because all the students use the same data available on the server, please don't alter the tables and the data. In task 3, you need to alter the emails table, hence every student should create an own emails table with the name of one of the group members as a prefix for the table name.

Exercise 2: Extensible Record Stores

DATA IN EXTENSIBLE RECORD STORES

Extensible record stores have tables as their basic data structure with a highly flexible column management. They implement the concept of column families that act as containers for subsets of columns.

A keyspaces in an Extensible Record Store is an object that holds together all column families of a design. It is the outermost container of the data in the data store. It resembles the schema concept in relational database management systems. Generally, there is one keyspaces per application. The tables and attributes of Enron Keyspace are shown in Figure 2.

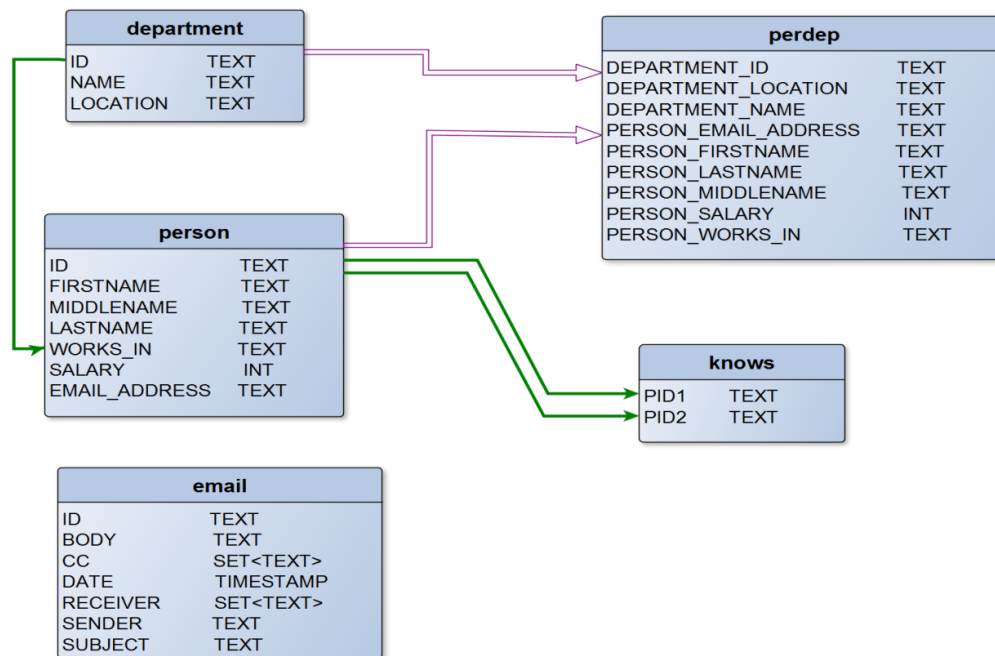


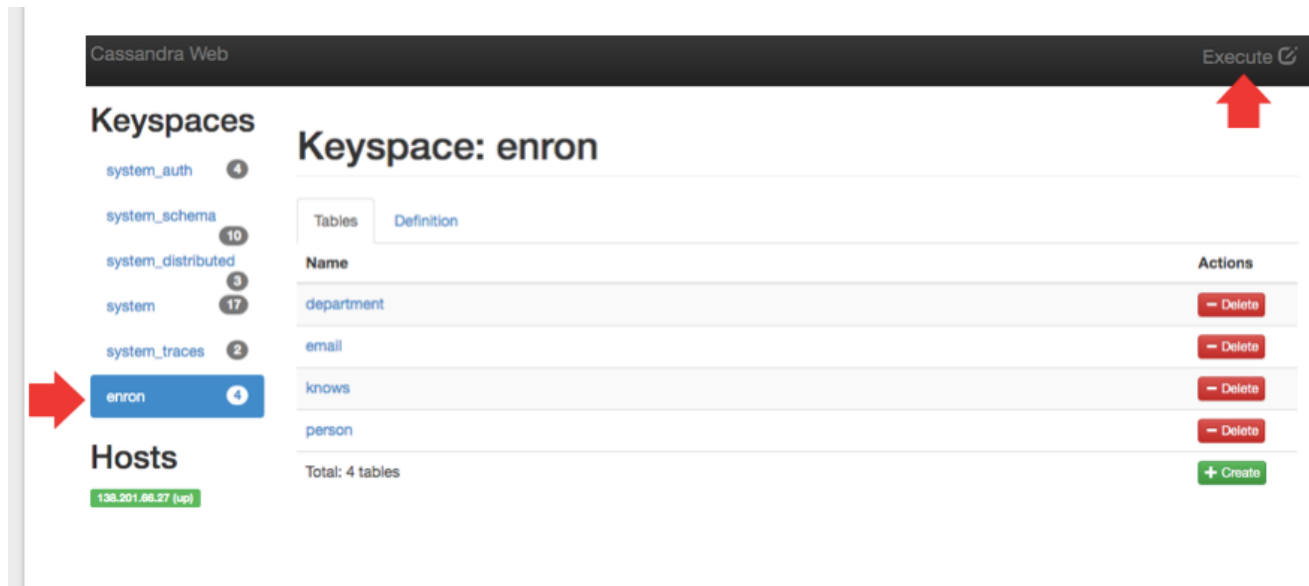
Figure 2. Enron keyspace in Cassandra container

System preparation

For extensible record store we use Cassandra server. There are two interface to the server:

1- Web GUI interface: you may use it to browse the data and run select queries. However, it does not run other type of queries like create, alter ... For those queries you need the CQLSH interface.

You may access the web GUI by opening the following page: <http://162.55.237.41:3000/enron>



You can see enron keyspace. Now you can run CQL queries against this keyspace by clicking on Execute. You can try some simple queries like:

- SELECT * FROM enron.department;
- SELECT * FROM enron.email;

2- **CQLSH interface** provides you with the possibility to run query on the shell of Cassandra. You need it in the tasks that require copying or altering the data or tables structure, or creating functions.

You may access it on the following page: http://162.55.237.41/run_query.php

You need your user name and password that you have used on PostgreSQL.

Exercises of Cassandra

Exercise 2 Task 1 Use Case: Equi Join

1.1 List of people with their department Max Time: 1h

For each person you want to know in which department she or he works. Therefore you have to make an output that contains a person's first name and last name and the name of the department she or he is working at.

1.2 Number of emails sent out per department Max Time: 2h

For each department: Find out how many emails in total were sent out from employees working there. The output per department shall contain the corresponding number of emails.

1.3 Number of emails received per department Max Time: 1

For each department: Find out how many emails in total were sent to employees working there (hint: carbon copies included). The output shall have the same structure as the output of Task 1.2.

Exercise 2 Task 2 Use Case: Theta Join

2.1 Correlation between salary and number of emails Max Time: 1h

Do people that earn more than the average salary in their department write more emails than those who don't?

Query for people that earn more than the average salary at their department and find out whether they write more emails than the other employees that earn less than the average salary at their department (equal is not considered). Check that for each department. First compute the result for the average salary (avg. S.) per department. Then produce the output of people that earn more than the avg. S. and accordingly produce the output for people who earn less than the average salary.

Produce a query result per department that contains the number of emails written by the people earning more and the people earning less than the average.

Exercise 2 Task 3 Use Case: Schema Evolution

3.1 Create a copy of a table Max Time: 0,5h

Create a copy of email table and name it with "email_yourname" (e.g. email_elmamooz) and execute the queries of task 3 (3.2, 3.3, 3.4) on this new table.

3.2 Add information for entity set Max Time: 0.5h

You have to introduce a new element (attribute) to the person's entity set (to your own copy of email table you created in task 3.1). Find the general syntax to do that.

3.3 Add information for entity set with default value Max Time: 0.5h

Now, use the syntax from *task 3.2* and add a new element "priority" to the person's entity set (your copy of table) with a default value of 1 for each entry. Then take a single entry of your choice (with a certain id) and set its priority to a value of 3.

3.4 drop a table Max Time: 0,5h

Drop the table you created in task 3.1.

Exercise 2 Task 4 Use Case: Missing values

4.1 Find missing values Max Time: 0.5h

Find missing values for each attribute of the e-mails. Which attribute has the most missing values?

Exercise 2 Task 5 Use Case: Range queries

5.1 Emails between two dates Max Time: 0.5h

Select all emails that have been written between the 01.09.2001 and the 31.10.2001. First, find out which date and time format is used in email!

5.2 Emails between two dates for Larry John May Max Time: 0.5h

Larry May is an employee of Enron. Find all emails he received between the 1st of September 2001 and the 31st of October 2001.

Exercise 2 Task 6 Use Case: Network analysis

Network analysis can be done to investigate social structures. In the related field of social network analysis a network is characterized by *nodes* (individual actors, people, or things within the network) and the *edges* (relationships or interactions) that connect them. To find out, how far from each other two nodes of the network are, we can count hops. To calculate the number of hops, we calculate $m+1$, where m is the number of intermediate nodes between the two nodes we're looking at. In Enron example, each employee is a node in the network and edges are represented by the emails sent from one to another employee or by the "knows" relationship between two employees.

6.1 Network size by e-mail Max Time: 1h

If the network nodes (the persons) are fully connected, how many hops are needed to reach everyone in Enron from Larry May by email? Consider the "from" and "to" fields to compute the amount of hops that is needed to reach everyone in Enron.

6.2 Network size by "knows" relation Max Time: 0.5h

How many hops are needed to reach everyone by their „knows" relationship (similar to task 6.1)?

6.3 2-hop email network Max Time: 0.5h

Which people are in the 2-hop email network? Again, consider the "knows" relationship, but only for people that are reachable with two hops.

6.4 Count outgoing edges Max Time: 0.5h

Find out who sent emails to exact 7 TO-recipients. The output shall contain the name(s) of the sender(s).

Exercise 2 Task 7 Use Case: User defined functions

7.1 Words and number of occurrences in a certain email Max Time: 4h

Take an email text and count the occurrence of each word in that email. The output shall contain the words and the number of occurrences. You might need to integrate Java code into an UDF (User defined function) to split the email text into words. Apply the UDF to a particular email body.

Hint: You might need an UDA (User defined aggregation) to produce the required output.

Bonus: Show the result according to

- the alphabetical order of the words
- the ascending order of occurrence
- the descending order of occurrence

7.2 Words and number of occurrences in all emails Max Time: 1.5h

Now, create an output similar to *task 7.1* for all emails. Info: depending on the database management system, this operation might be very expensive and can result in a timeout. In that case you might want to set a threshold for a timeout, if possible.