CSCI 5408 (Winter 2017)

Assignment 2 Tutorial

Abhinav Kalra

25 January,2017

DALHOUSIE UNIVERSITY
Inspiring Minds

Agenda

- To learn concepts of Distributed Database Systems running on Clouds
- To learn load balancing and scalability on mission critical applications
- To learn fault tolerance and performance tuning using Clouds
- To learn Cassandra database system and firewall concepts on Clouds

Centralized Database Systems

- Data stored centrally
- DBMS processes queries originating from multiple remote sites
- Easy to store, organize and manage data

Disadvantages of Centralized Databases

- Ever growing volume of data
- Difficult to scale
- Cant effectively cope with high request traffic
- Limited fault tolerance
- Limited Data redundancy

Distributed Databases

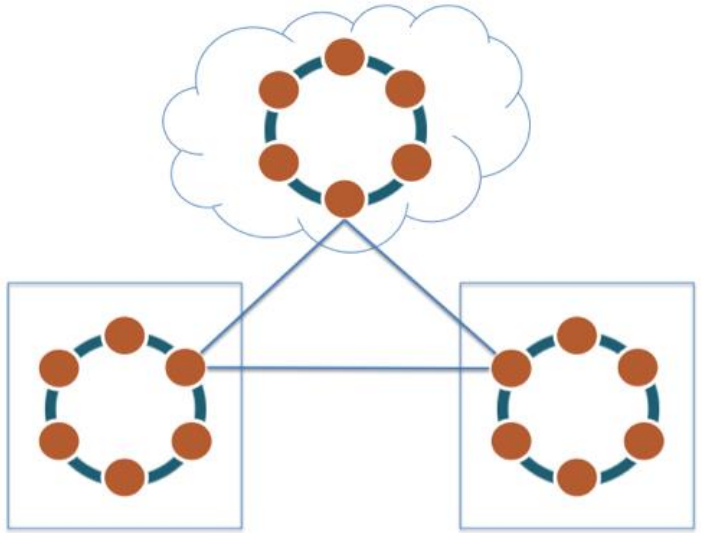
- Set of interconnected sites which host a Database system
- Communicate over a network
- Logically interrelated but physically separate systems
- Managed by Distributed DBMS
- Data split into fragments which may be replicated at multiple sites
- Increases data redundancy

Cassandra

- Distributed and decentralized Database system
- Originated at Facebook and subsequently open sourced
- Highly Scalable with no Single Point of Failure
- Linear scalability
- Uses Cassandra Query Language similar to SQL !!

Cassandra

- Master-less Ring architecture
- No concept of master node
- Data replicated across one or more nodes in cluster
- Nodes communicate via Gossip Protocol
- Cluster may encompass several datacenters
- Data distributed over several data centers



"A Brief Introduction To Apache Cassandra | Datastax Academy: Free Cassandra Tutorials And Training". *Academy.datastax.com*.

Cassandra

- Column-family Data model
- *Keyspace:* stores several Column Families
- *Column-Family:* Container for collection of rows
- *Columns:* can be dynamically added to column families
- *Row:* a unit of replication in Cassandra
- Fetch data as an ordered key-value pair comprising *(row,column-column-value)*

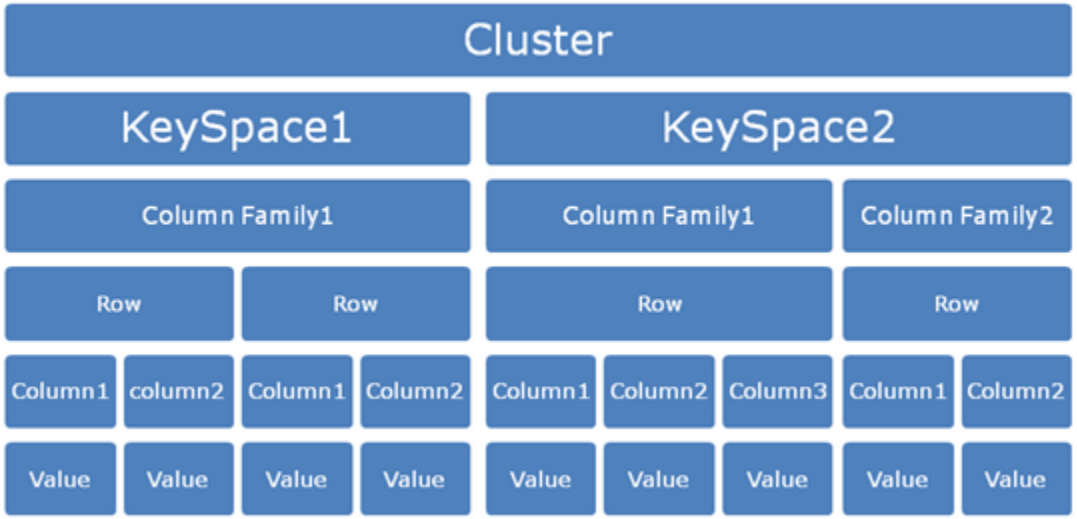


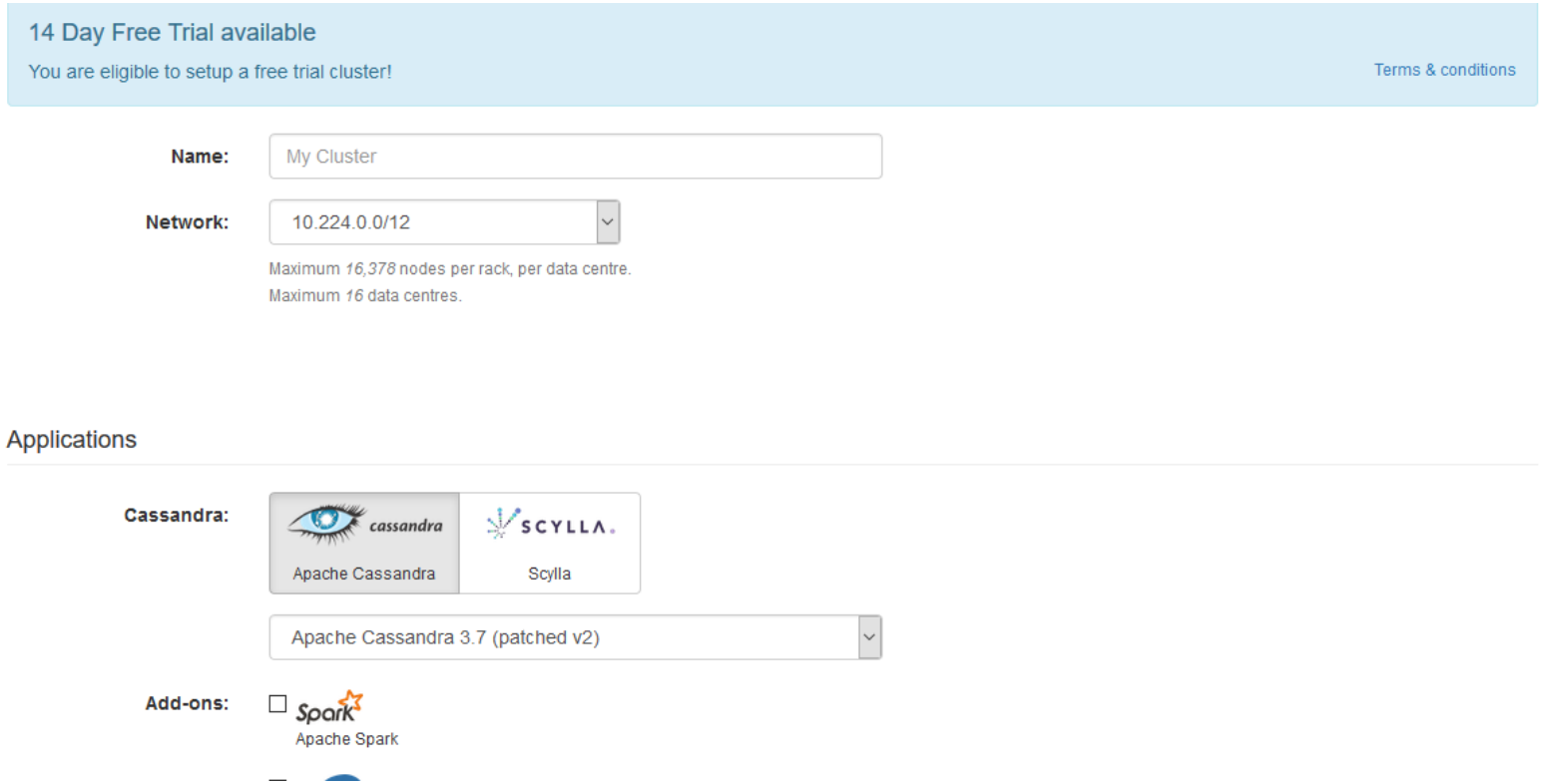
Image courtesy: Karthikeyan, K, K Karthikeyan."Understanding NOSQL Through Cassandra II". *Ksivakarthikeyan.blogspot.ca*

Installing Cassandra Cluster

- We are going to install a managed Cassandra cluster in Instaclustr which has a 14 day trial period
- Sign up for the free account at : https://console.instaclustr.com/user/signup

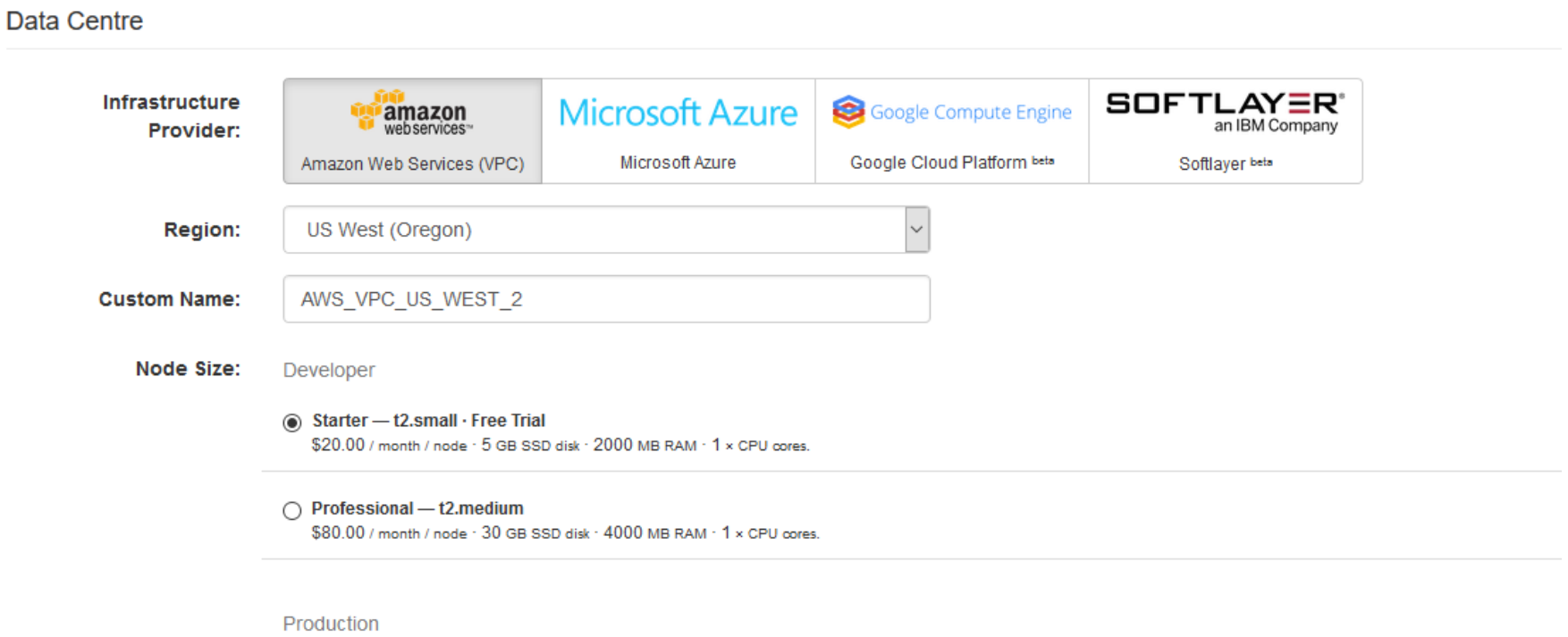
Installing Cassandra Cluster

- Once the account is created, login and create new Cassandra Cluster



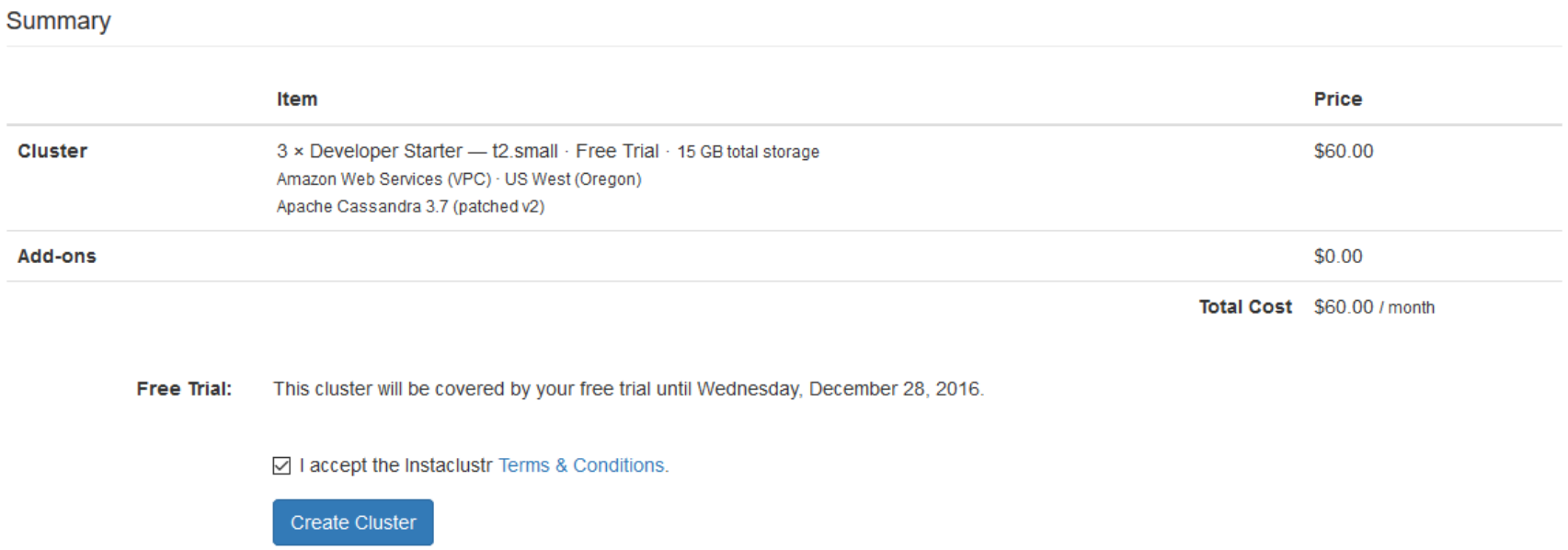
Installing Cassandra Cluster

- Choose Data Centre



Installing Cassandra Cluster

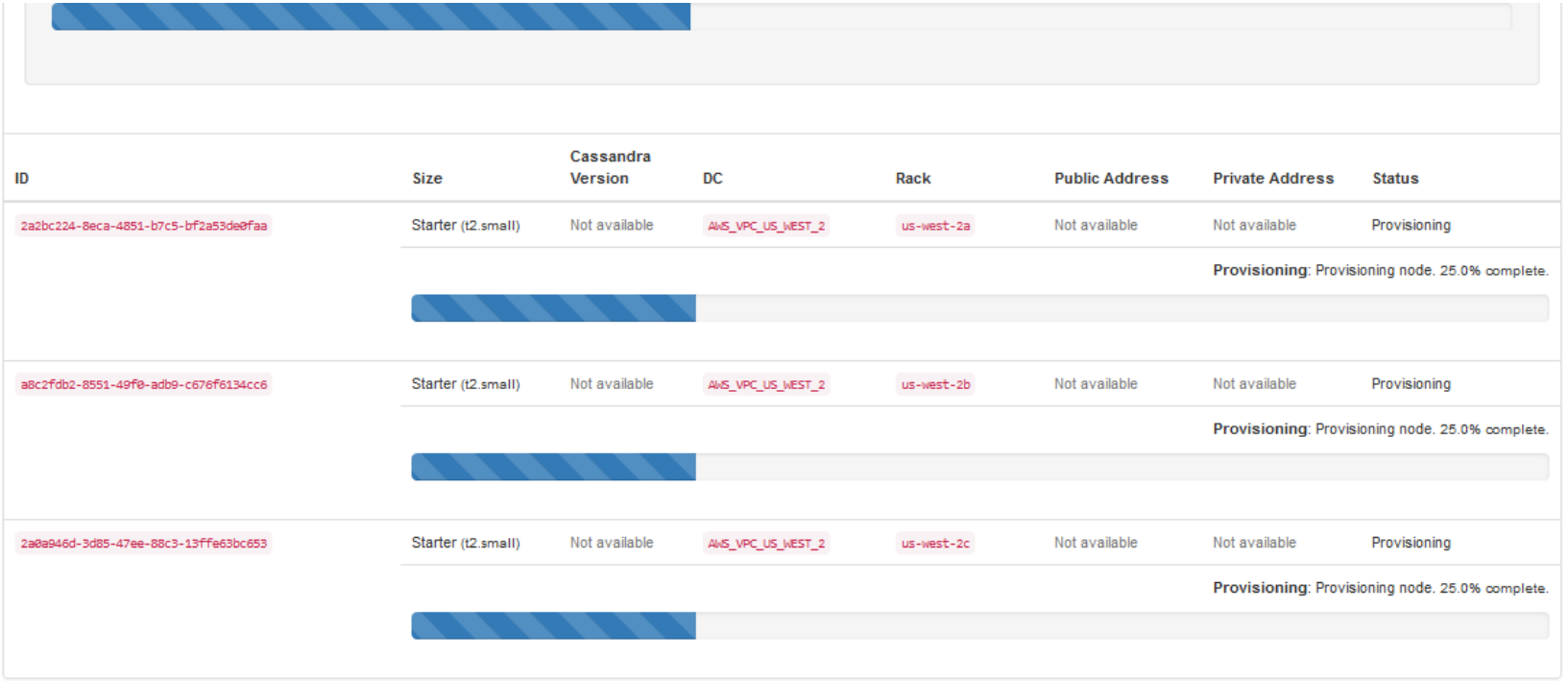
- Create Cluster



DALHOUSIE UNIVERSITY
Inspiring Minds

Installing Cassandra Cluster

- Provisioning Cassandra Cluster

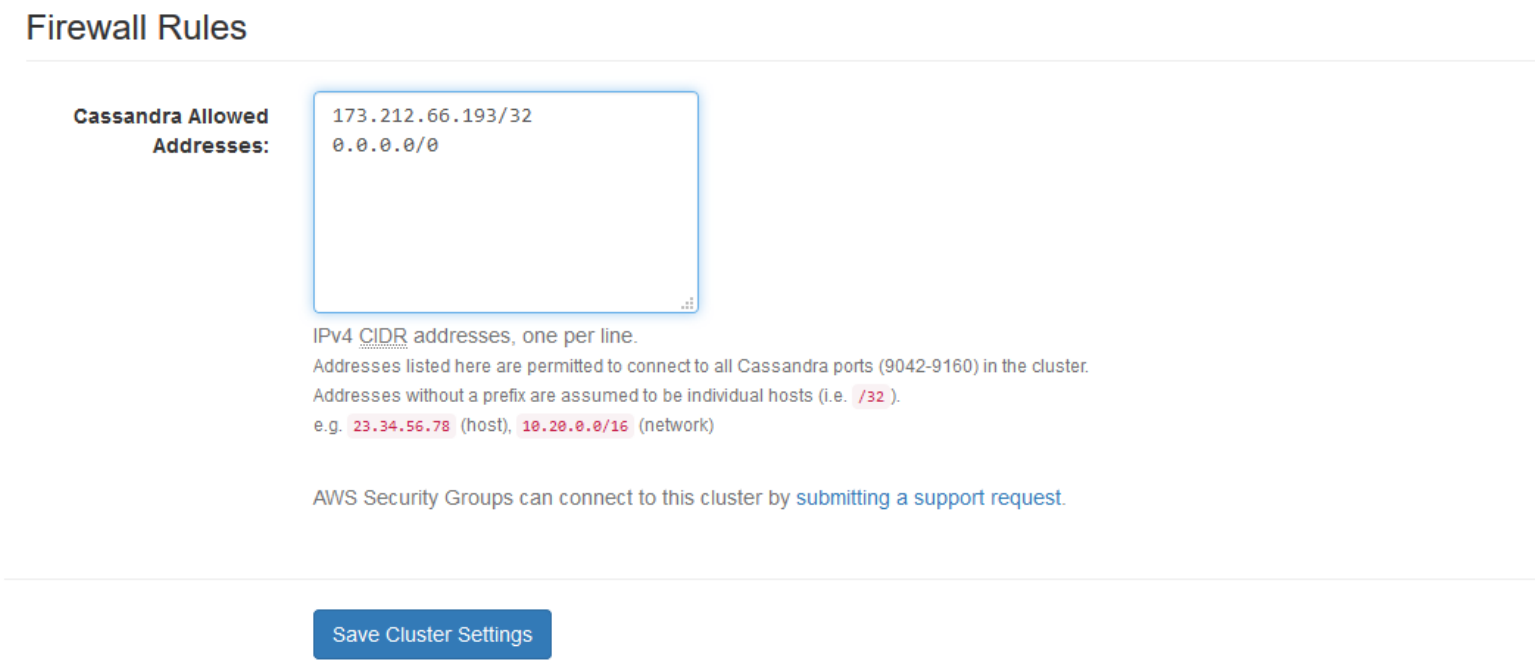


ID	Size	Cassandra Version	DC	Rack	Public Address	Private Address	Status
2a2bc224-8eca-4851-b7c5-bf2a53de0faa	Starter (t2.small)	Not available	AWS_VPC_US_WEST_2	us-west-2a	Not available	Not available	Provisioning
						Provisioning: Provisioning node. 25.0% complete.	
a8c2fdb2-8551-49f0-adb9-c676f6134cc6	Starter (t2.small)	Not available	AWS_VPC_US_WEST_2	us-west-2b	Not available	Not available	Provisioning
						Provisioning: Provisioning node. 25.0% complete.	
2a0a946d-3d85-47ee-88c3-13ffe63bc653	Starter (t2.small)	Not available	AWS_VPC_US_WEST_2	us-west-2c	Not available	Not available	Provisioning
						Provisioning: Provisioning node. 25.0% complete.	

DALHOUSIE UNIVERSITY
Inspiring Minds

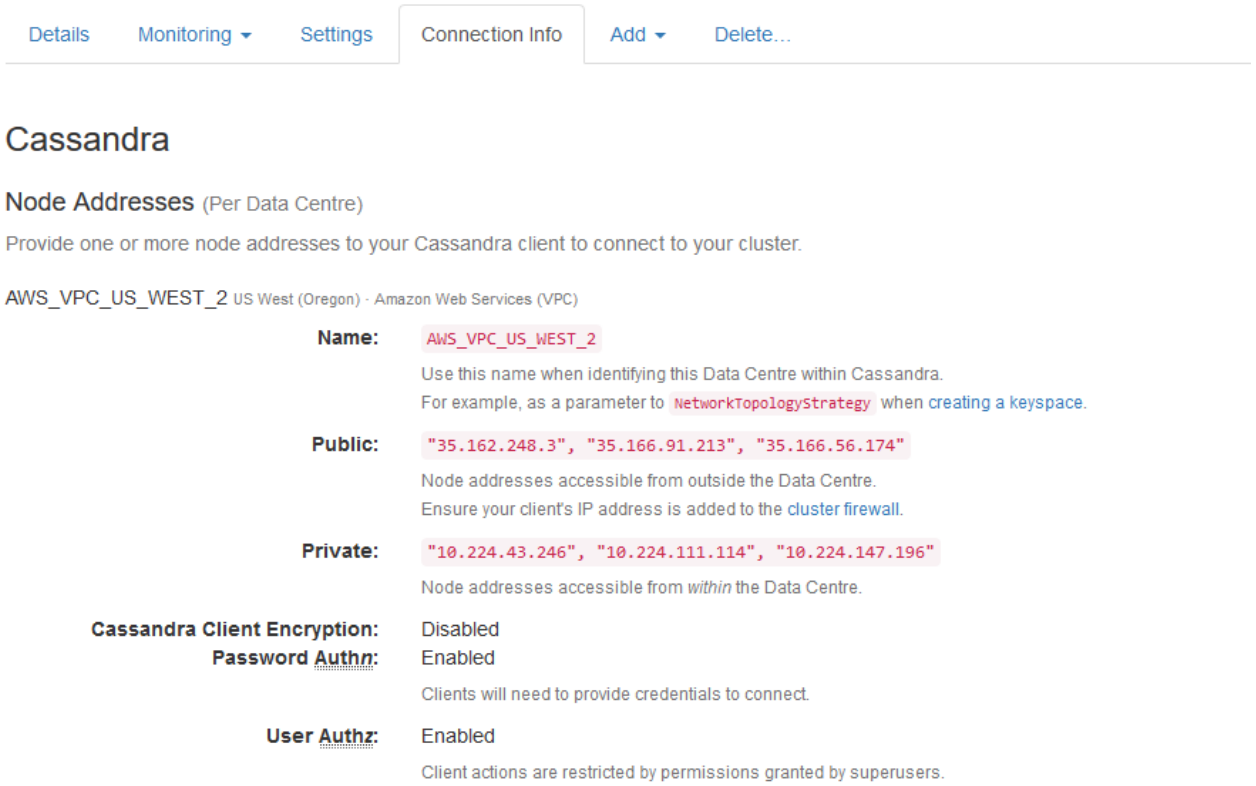
Installing Cassandra Cluster

- Configure Firewall rules



Installing Cassandra Cluster

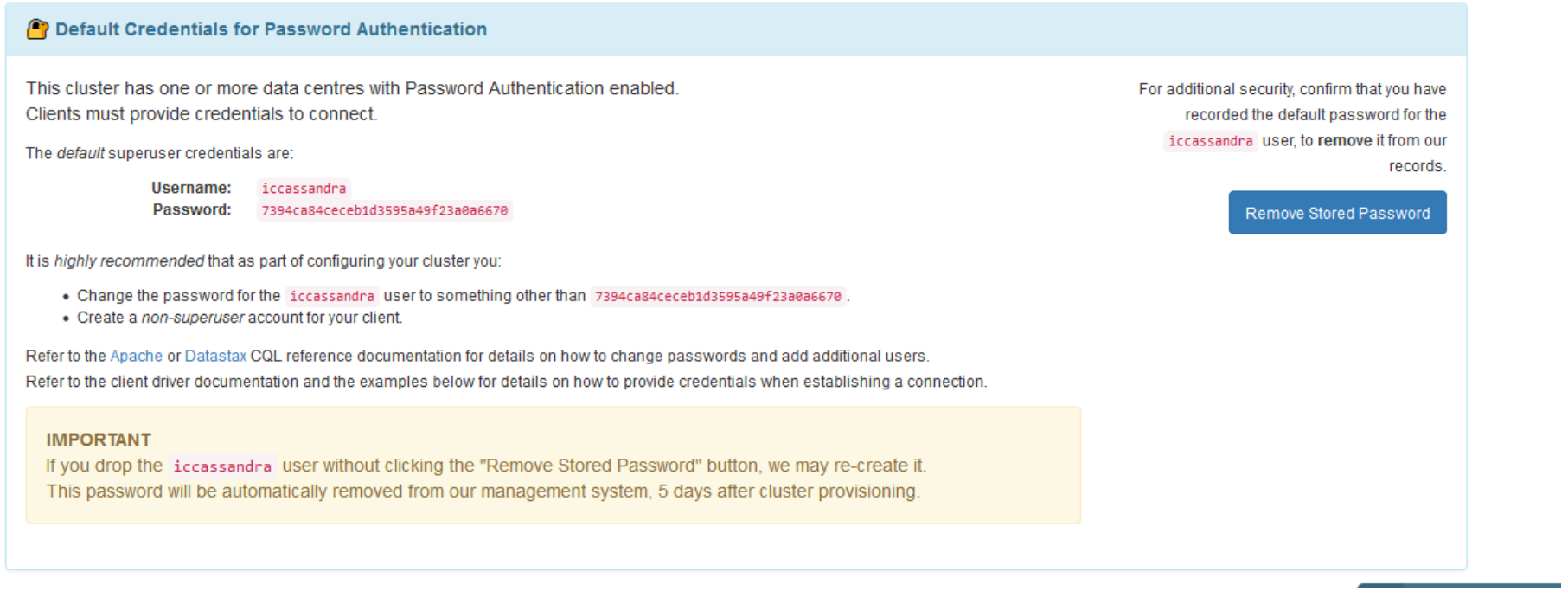
- Node Addresses



DALHOUSIE UNIVERSITY
Inspiring Minds

Installing Cassandra Cluster

- Credentials for authentication



DALHOUSIE UNIVERSITY
Inspiring Minds

Installing Cassandra Cluster

- Cluster builder and connection strings

cqlsh | Java | Python | Ruby | Other

For more information on connecting to Cassandra with the DataStax Cassandra Driver for Java refer to:

- DataStax Java Driver for Apache Cassandra (GitHub)
- DataStax Java Driver for Apache Cassandra - 3.0 API(JavaDoc)
- Java Driver 3.0 for Apache Cassandra (Reference Manual)

```
final Cluster.Builder clusterBuilder = Cluster.builder()
    .addContactPoints(
        "35.162.248.3", "35.166.91.213", "35.166.56.174" // AWS_VPC_US_WEST_2 (Amazon Web Services (VPC))
    )
    .withLoadBalancingPolicy(DCAwareRoundRobinPolicy.builder().withLocalDc("AWS_VPC_US_WEST_2").build()) // your local data centre
    .withPort(9042)
    .withAuthProvider(new PlainTextAuthProvider("iccassandra", "7394ca84ceceb1d3595a49f23a0a6670"));

try (final Cluster cluster = clusterBuilder.build()) {
    final Metadata metadata = cluster.getMetadata();
    System.out.printf("Connected to cluster: %s\n", metadata.getClusterName());

    for (final Host host: metadata.getAllHosts()) {
        System.out.printf("Datacenter: %s; Host: %s; Rack: %s\n", host.getDatacenter(), host.getAddress(), host.getRack());
    }
}
```

Connecting to Cassandra

- CQLSH utility to connect to Cassandra and upload data comes as part of Cassandra distribution
- The Cassandra distribution can be downloaded from *http://cassandra.apache.org/download/*
- The command to connect to Cassandra via CQLSH is
  - cqlsh public_ip_of_your_node 9042 –u your_username -p your_password

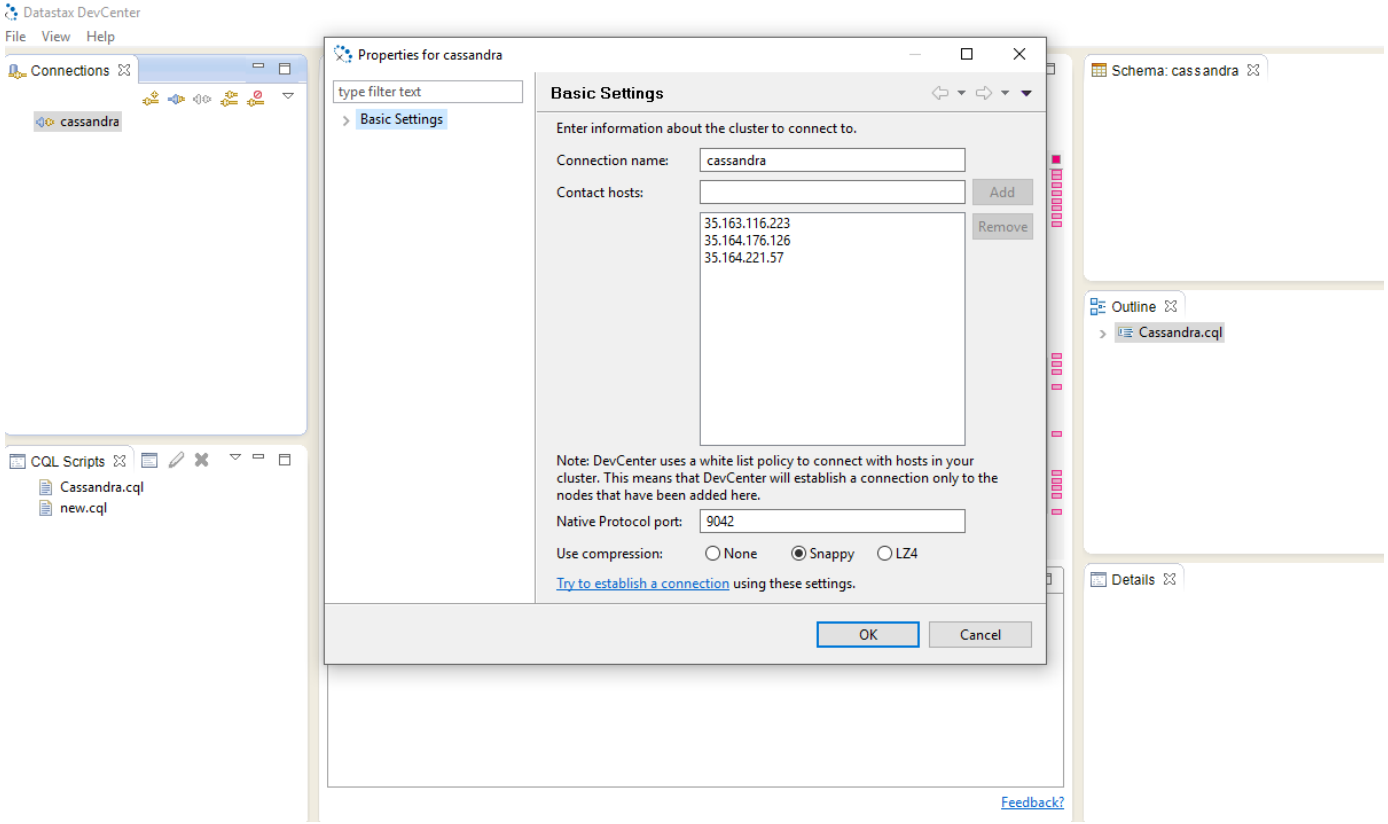
```
C:\Program Files\DataStax-DDC\apache-cassandra\bin>cqlsh.bat  35.162.248.3 9042 -u iccassandra -p 7394ca84ceceb1d3595a49f23a0a6670

WARNING: console codepage must be set to cp65001 to support utf-8 encoding on Windows platforms.
If you experience encoding problems, change your console codepage with 'chcp 65001' before starting cqlsh.

Connected to myCluster at 35.162.248.3:9042.
[cqlsh 5.0.1 | Cassandra 3.7.2 | CQL spec 3.4.2 | Native protocol v4]
Use HELP for help.
WARNING: pyreadline dependency missing.  Install to enable tab completion.
iccassandra@cqlsh> select * from iccassandra;
```

Querying Cassandra

- Cassandra Cluster can be queried from CQLSH.
- Alternatively, DataStax DevCenter is a graphical UI to query Cassandra
- It can be downloaded from: *https://www.datastax.com/products/datastax-devcenter-and-development-tools#DataStax-DevCenter*
- You can use either CQLSH or DevCenter to create your keyspace, table and try out your queries and learn.

Querying Cassandra

- Create a connection to your Cassandra Instance in DataStax DevCenter
- Provide the login credentials and Public IP address



DALHOUSIE UNIVERSITY
Inspiring Minds

Loading Data

- Once the key space and table is defined, import the Philadelphia crime data CSV file to Cassandra
- You may have to perform data cleaning and preprocessing
- Bulk copy from CQLSH : cqlsh> copy <keyspace>.<table> from <Directory>/<filename>.csv with delimiter=',';

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Dc_Dist	Psa	Dispatch_Date_Tir	Dispatch_Date	Dispatch_	Hour	Dc_Key	Location_	UCR_Gene	Text_Gen	Police_Di	Month	Lon	Lat
2	18	3	2009-10-02 14:24	2009-10-02	14:24:00	14	2.01E+11	S 38TH ST	800	Other Assaults		Oct-09		
3	14	1	2009-05-10 0:55	2009-05-10	0:55:00	0	2.01E+11	8500 BLOC	2600	All Other Offenses		May-09		
4	25	J	2009-08-07 15:40	2009-08-07	15:40:00	15	2.01E+11	6TH CAME	800	Other Assaults		Aug-09		
5	35	D	2009-07-19 1:09	2009-07-19	1:09:00	1	2.01E+11	5500 BLOC	1500	Weapon \	20	Jul-09	-75.1305	40.03639
6	9	R	2009-06-25 0:14	2009-06-25	0:14:00	0	2.01E+11	1800 BLOC	2600	All Other	8	Jun-09	-75.1664	39.96953
7	17	1	2015-04-25 12:50	2015-04-25	12:50:00	12	2.02E+11	800 BLOCK	600	Thefts	13	Apr-15	-75.1664	39.94007
8	23	K	2009-02-10 14:33	2009-02-10	14:33:00	14	2.01E+11	2200 BLOC	800	Other Ass	16	Feb-09	-75.1711	39.97959
9	77	A	2009-04-02 18:30	2009-04-02	18:30:00	18	2.01E+11	TERMINAL	500	Burglary Non-Reside		Apr-09		
10	35	D	2009-03-18 1:14	2009-03-18	1:14:00	1	2.01E+11	N 5TH SON	2600	All Other Offenses		Mar-09		
11	23	L	2009-06-14 20:30	2009-06-14	20:30:00	20	2.01E+11	N 2828 W	2600	All Other Offenses		Jun-09		
12	22	P	2009-01-19 16:44	2009-01-19	16:44:00	16	2.01E+11	N 2800 ST	400	Aggravated Assault F		Jan-09		
13	1	J	2009-02-09 22:52	2009-02-09	22:52:00	22	2.01E+11	2700 BLOC	800	Other Assaults		Feb-09		
14	22	3	2015-10-06 18:18	2015-10-06	18:18:00	18	2.02E+11	1500 BLOC	600	Thefts	16	Oct-15	-75.1602	39.97613
15	22	3	2015-10-09 0:49	2015-10-09	0:49:00	0	2.02E+11	1500 BLOC	600	Thefts	16	Oct-15	-75.1631	39.97787
16	77	A	2015-05-03 20:39	2015-05-03	20:39:00	20	2.02E+11	GATE C-18	600	Thefts		May-15		
17	2	1	2015-11-30 8:39	2015-11-30	8:39:00	8	2.02E+11	6500 BLOC	600	Thefts	2	Nov-15	-75.0756	40.04127

- Please refer to below two tutorials for bulk insertion of data:
  - http://code-mix.blogspot.ca/2013/08/import-data-from-csv-file-to-cassandra.html
  - http://prashantguptaimpetus.blogspot.ca/2013/02/bulk-insert-in-cassandra.html

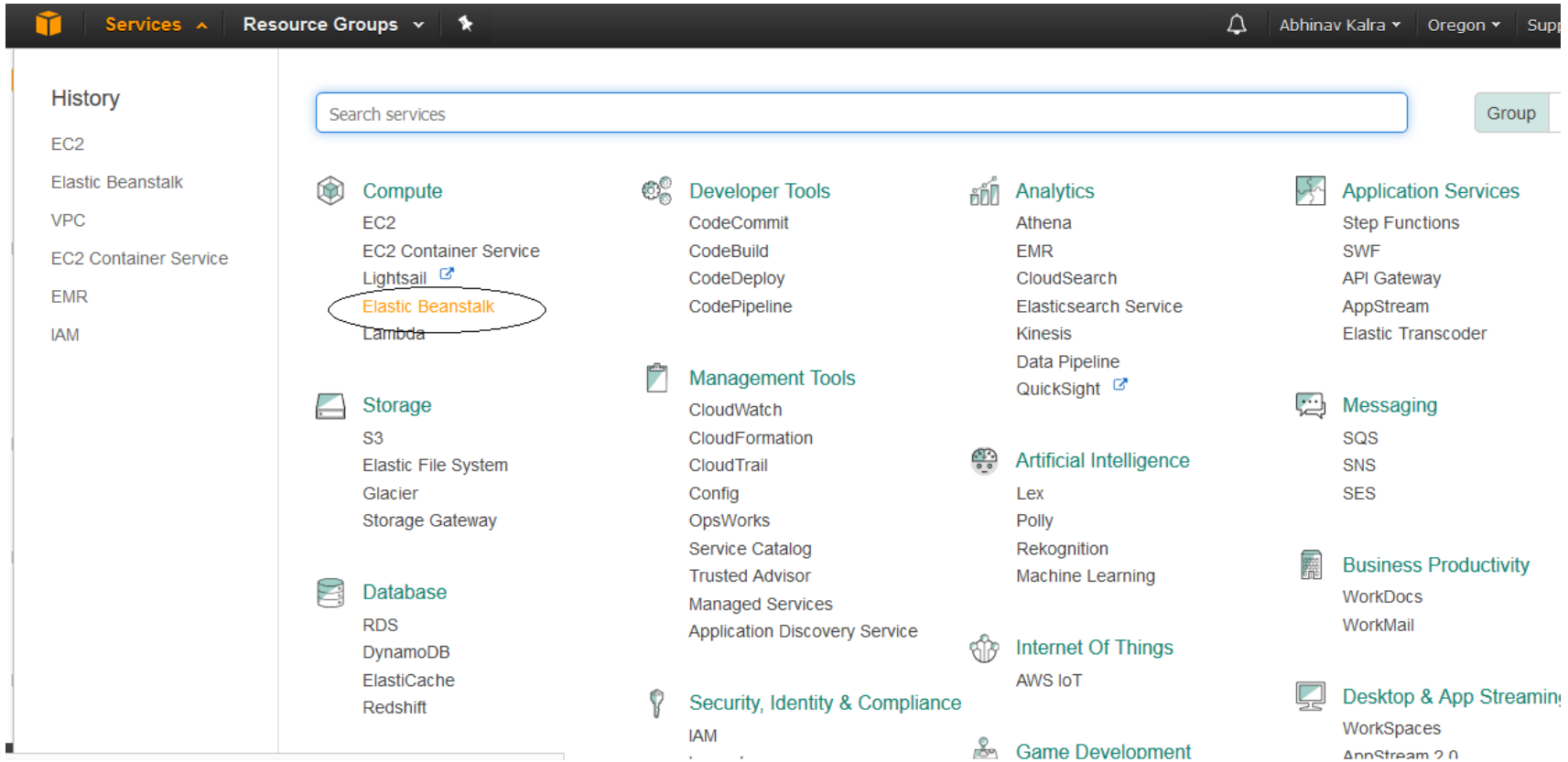
Query Data

- Cassandra uses Cassandra Query Language to query data
- The reference to CQL is available at:
  - *https://cassandra.apache.org/doc/latest/getting_started/index.html*
  - *http://docs.datastax.com/en/cql/3.1/cql/cql_intro_c.html*

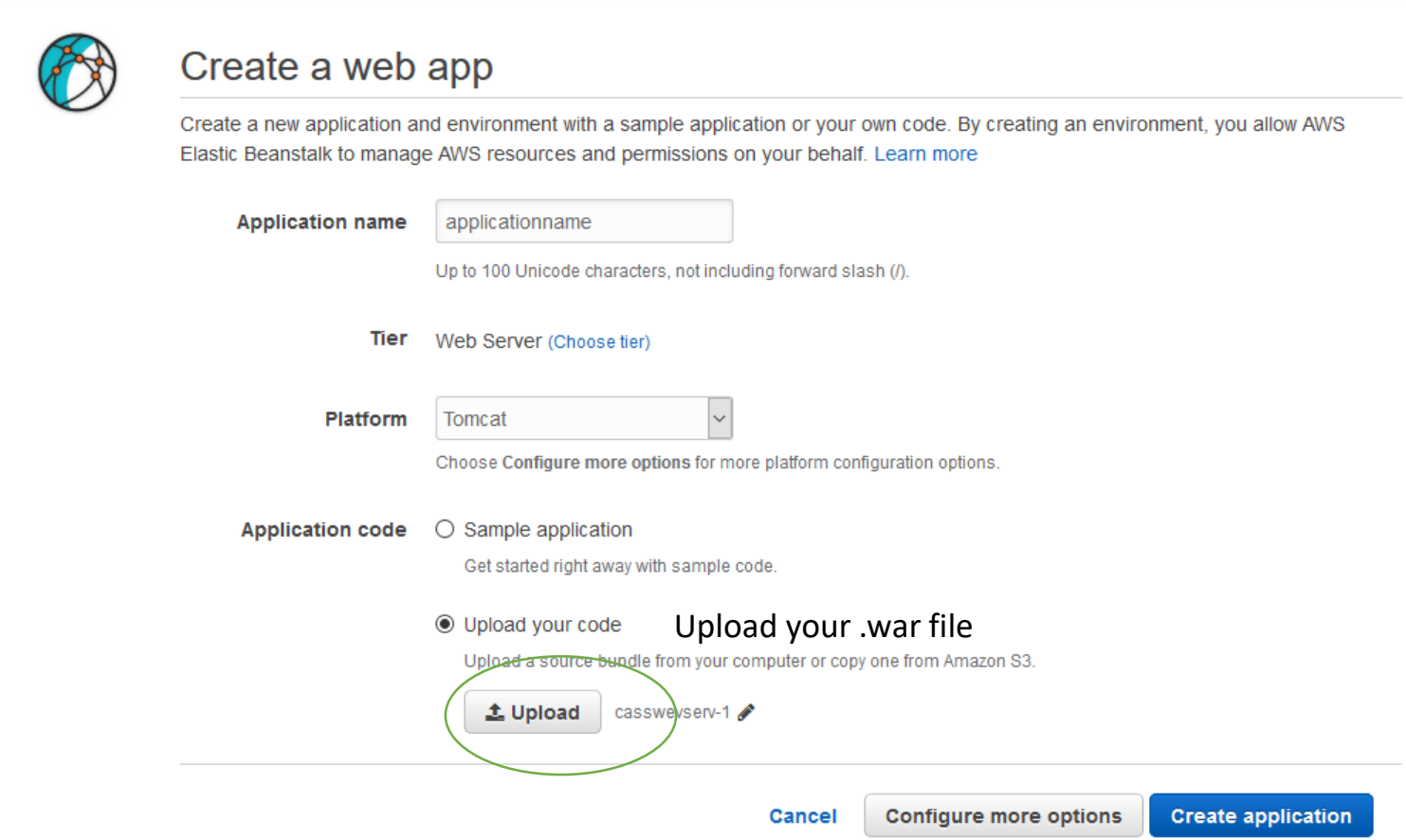
- For the UI application, drivers for Cassandra can be downloaded from Datastax or from link below:
  - *http://cassandra.apache.org/doc/latest/getting_started/drivers.html*
  - *http://docs.datastax.com/en/developer/driver-matrix/doc/common/driverMatrix.html*

Deploying Web Service

- Deploy your web service on AWS through Elastic Beanstalk

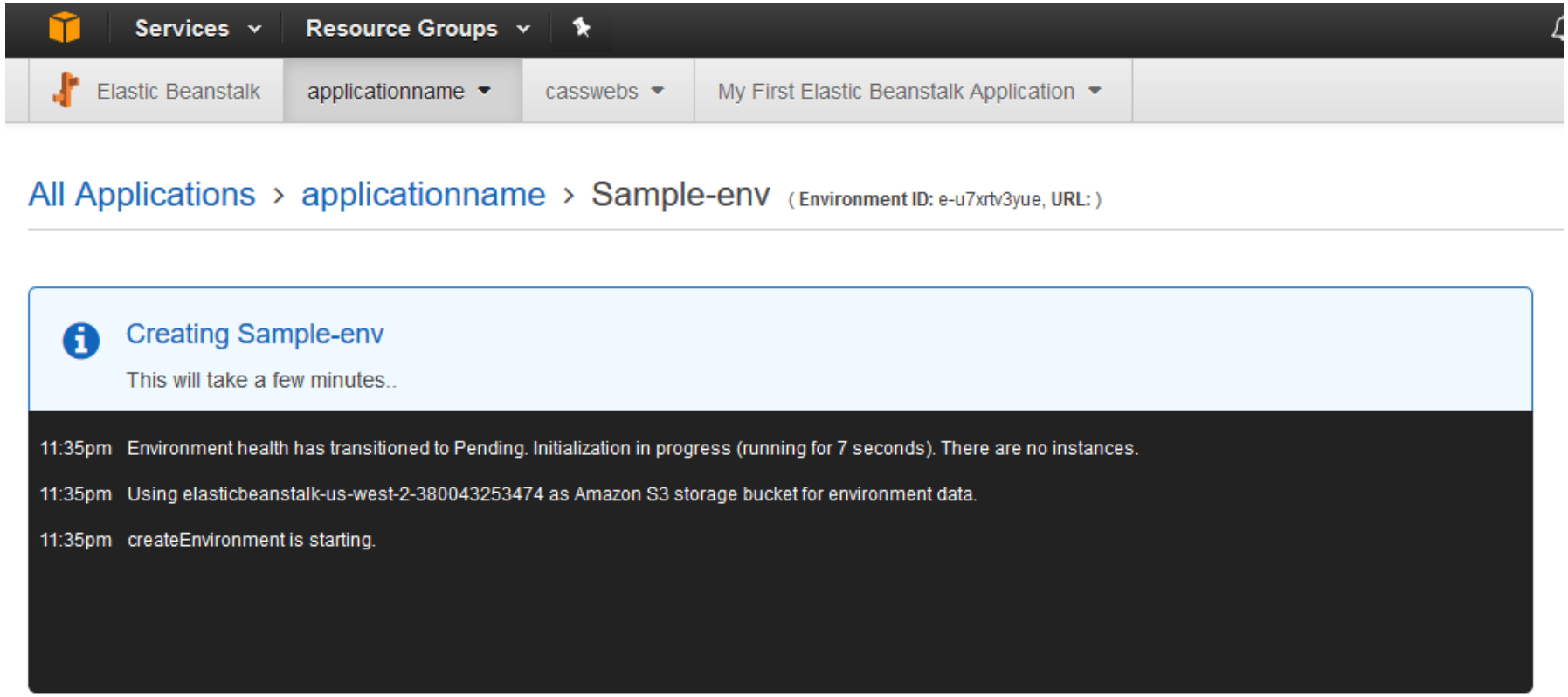


Deploying Web Service

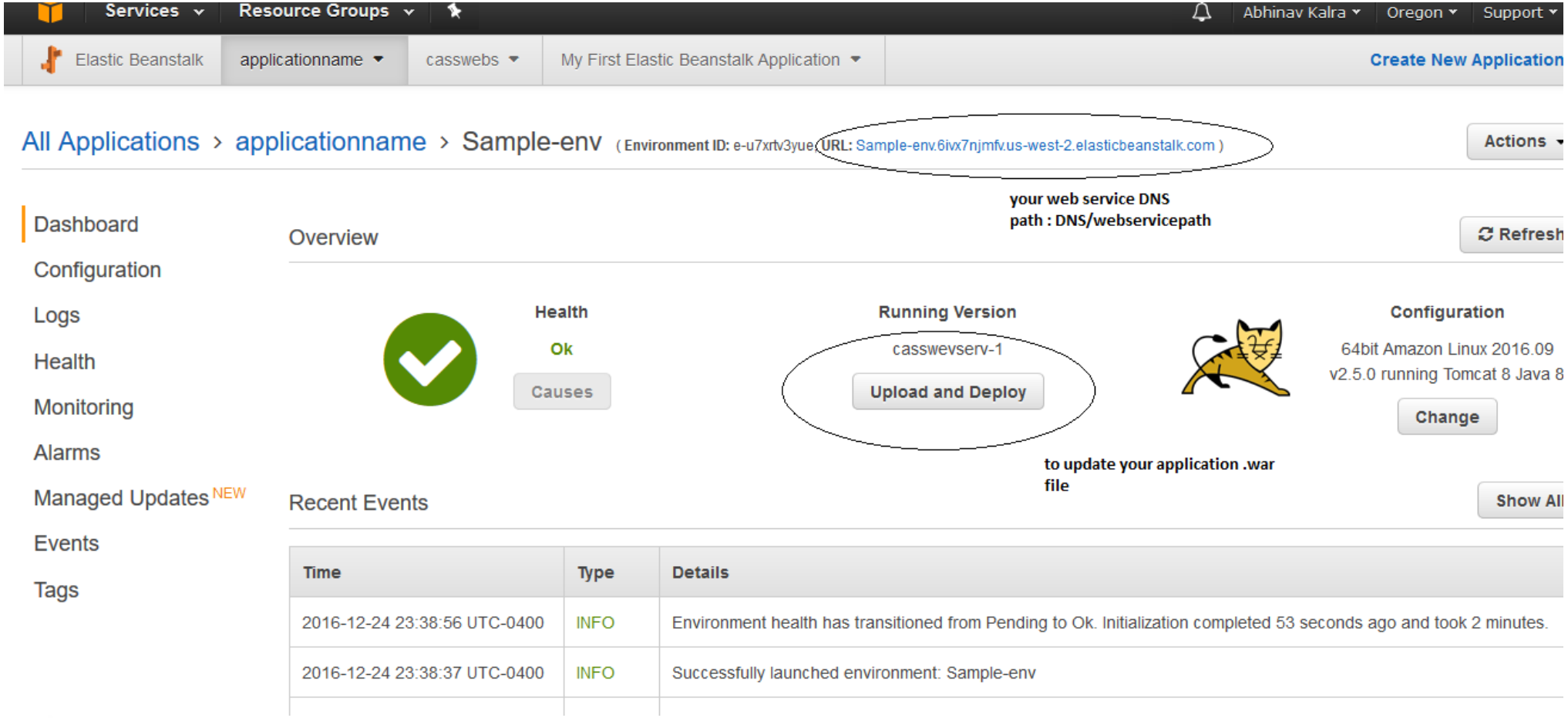


DALHOUSIE UNIVERSITY
Inspiring Minds

Deploying Web App



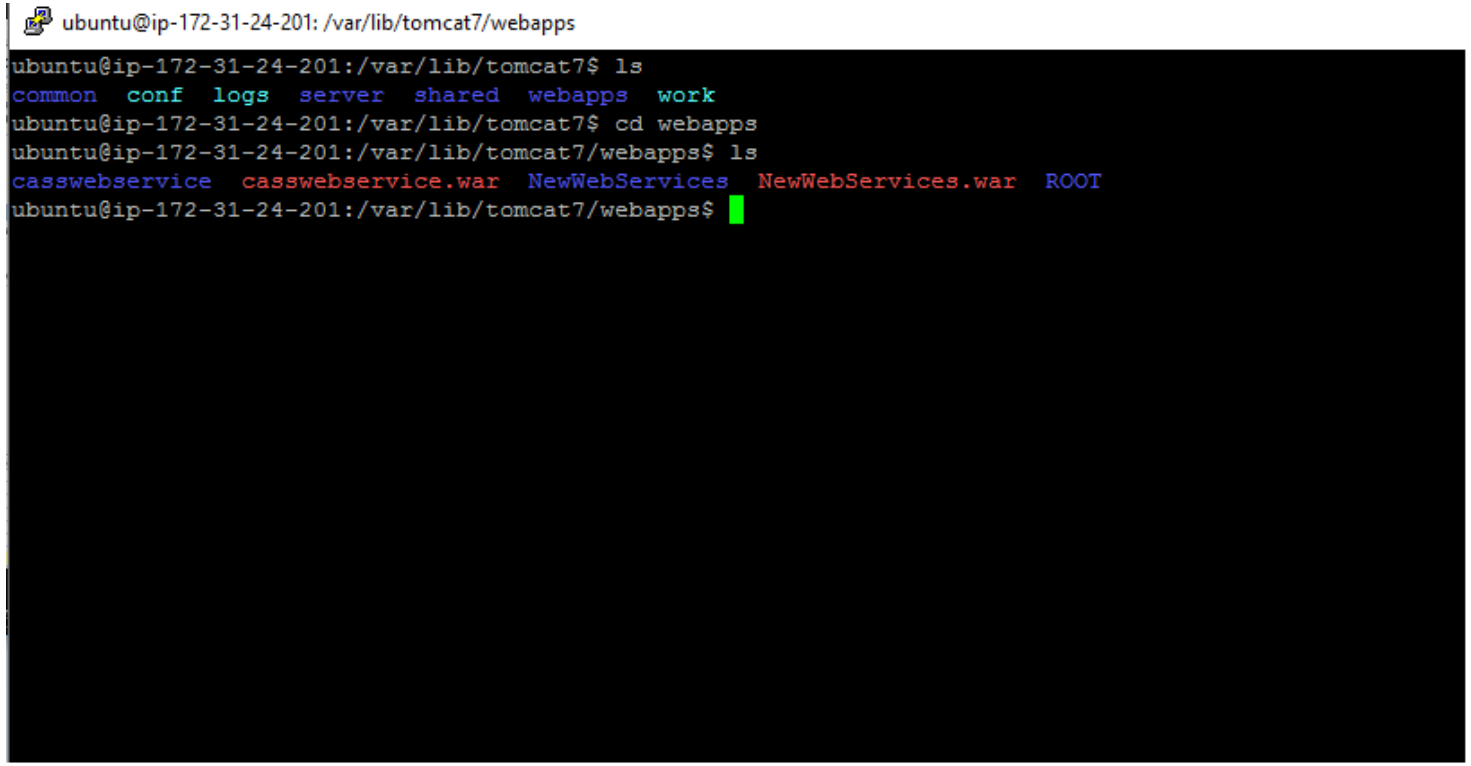
Deploying Web App



DALHOUSIE UNIVERSITY
Inspiring Minds

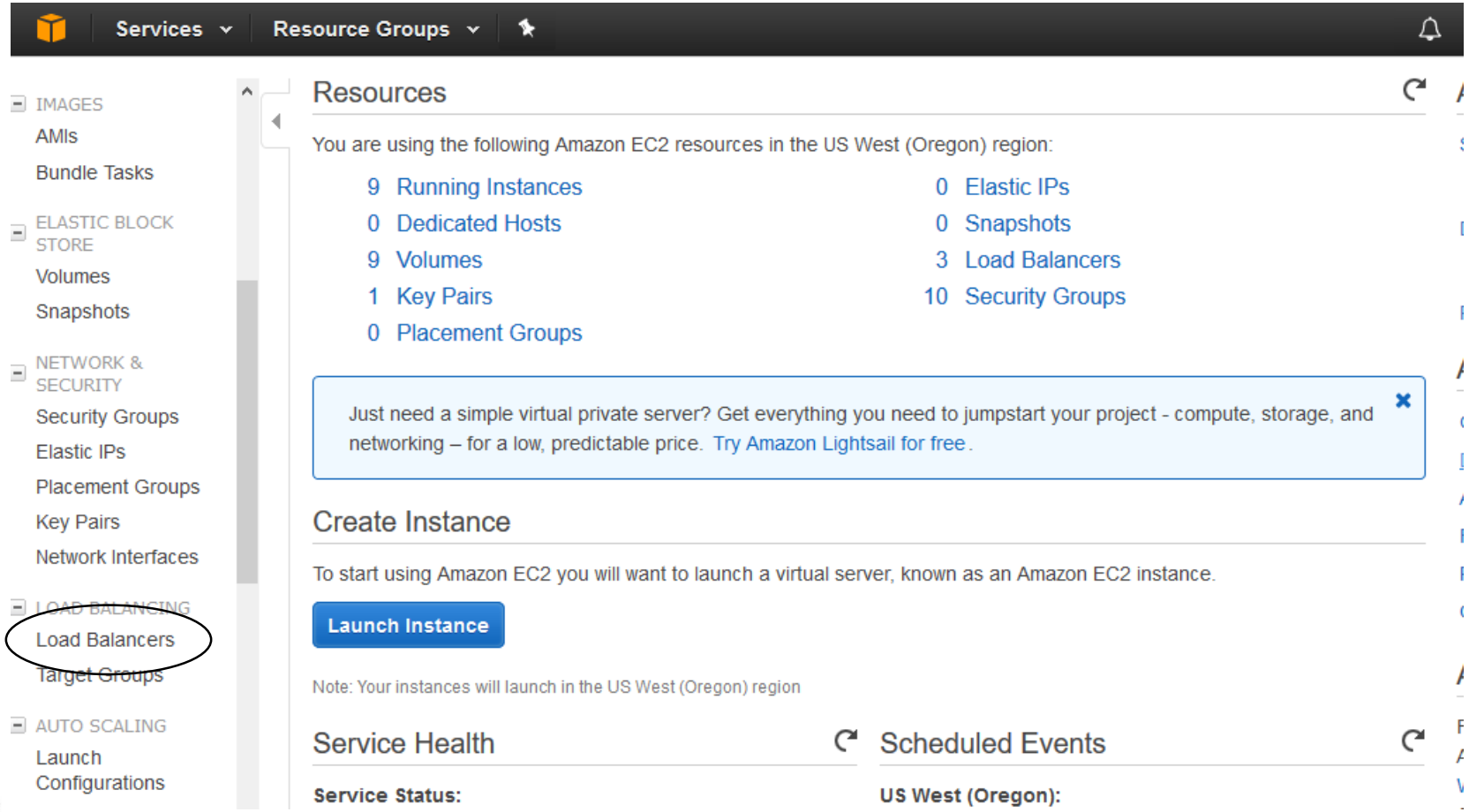
Deploying Web App

- Alternately, the .war file can be copied to the Tomcat server's *webapps* directory on your EC2 instance.
- For this, you should create a Free tier EC2 instance and install JRE and Tomcat7 on it.



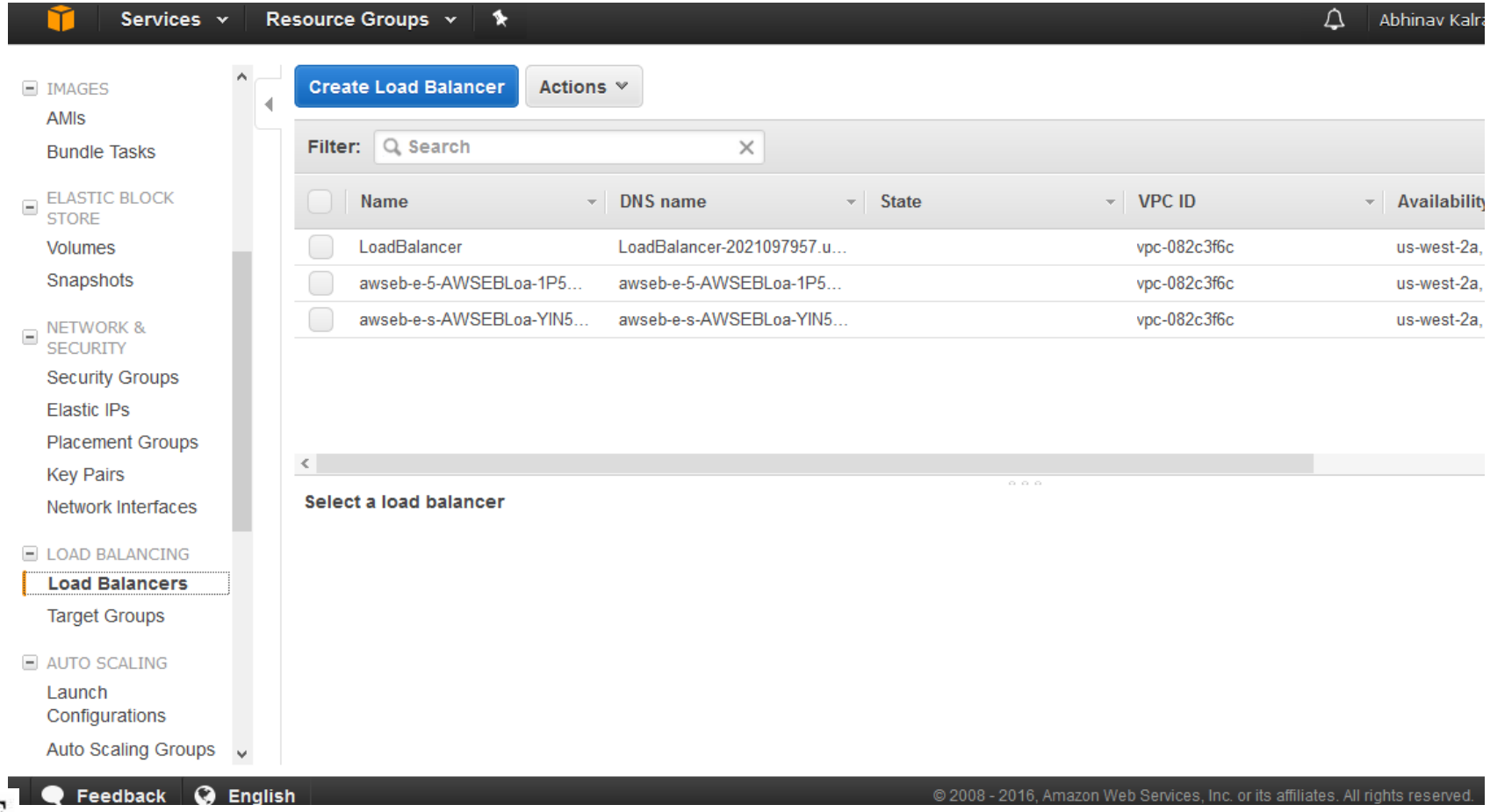
Creating Load Balancer on AWS

* Load Balancer service incurs a small charge



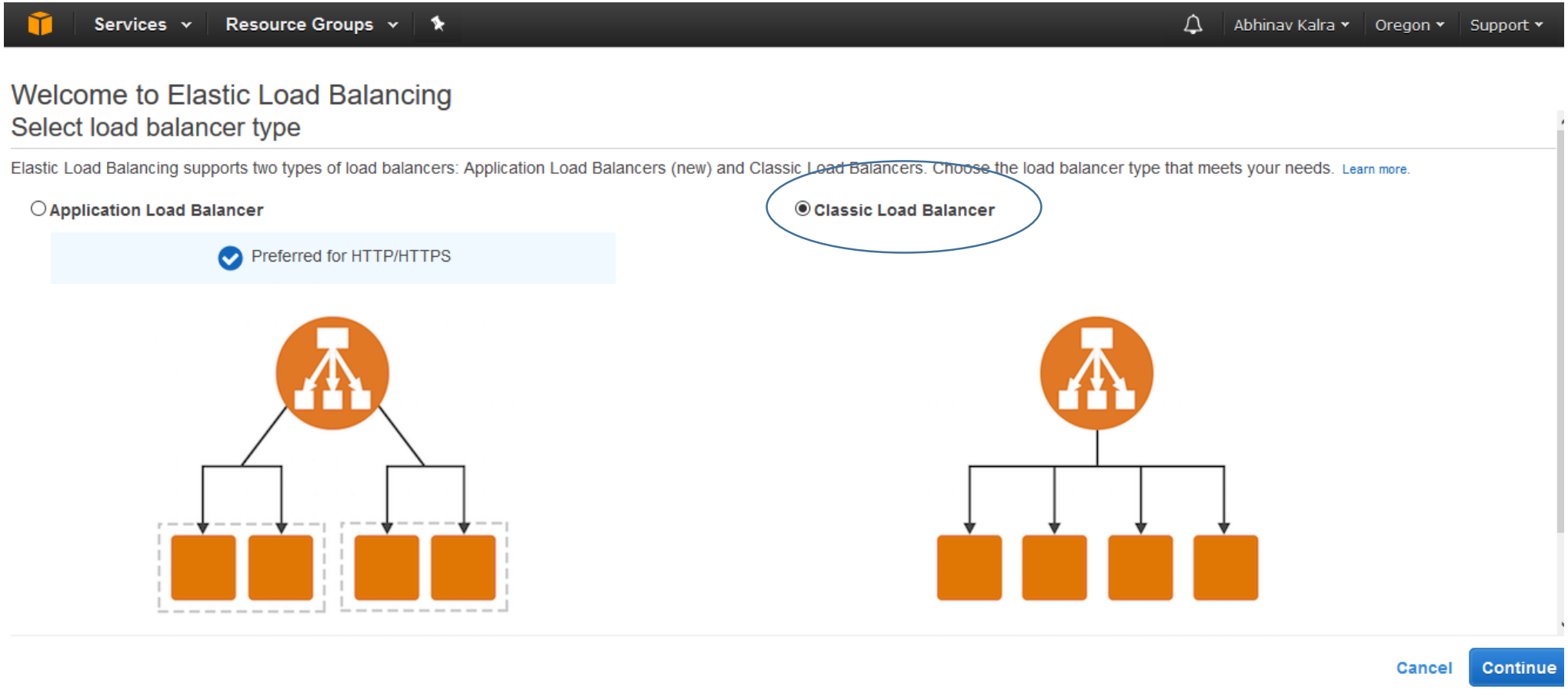
DALHOUSIE UNIVERSITY
Inspiring Minds

Creating Load Balancer on AWS

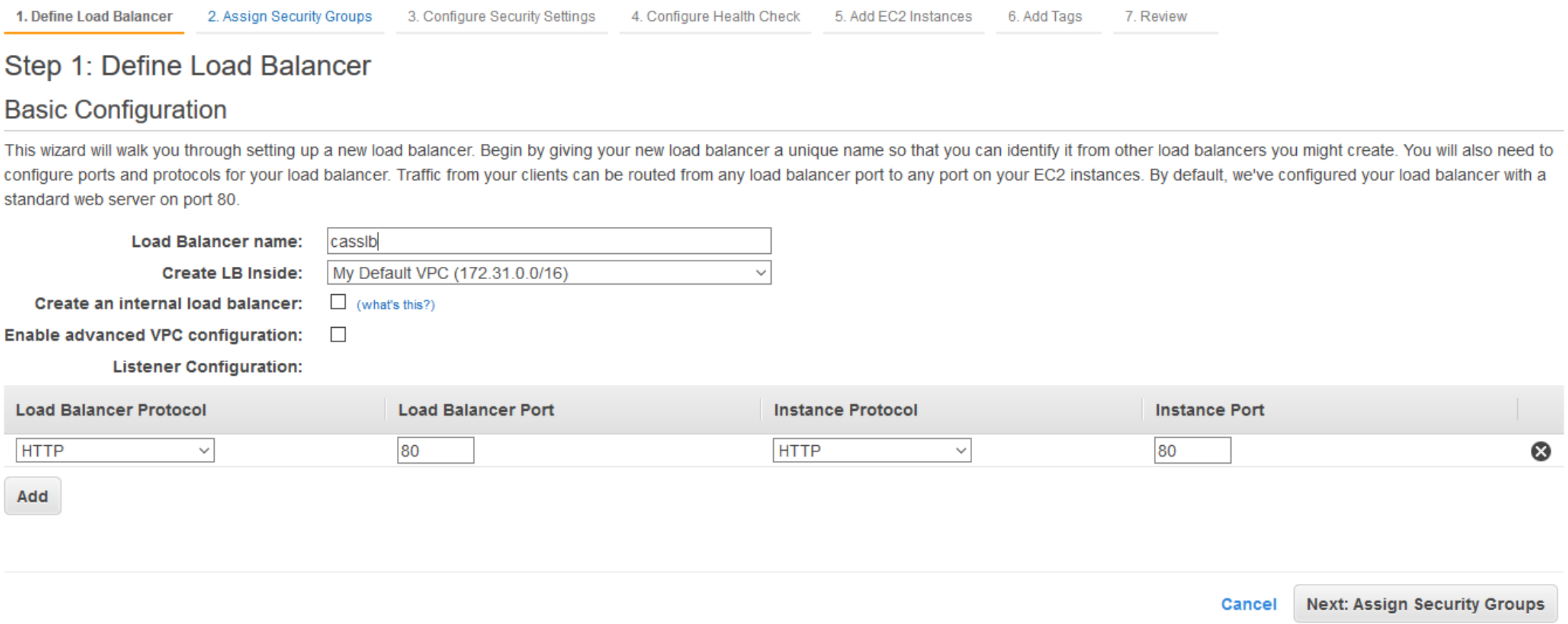


DALHOUSIE UNIVERSITY
Inspiring Minds

Creating Load Balancer on AWS

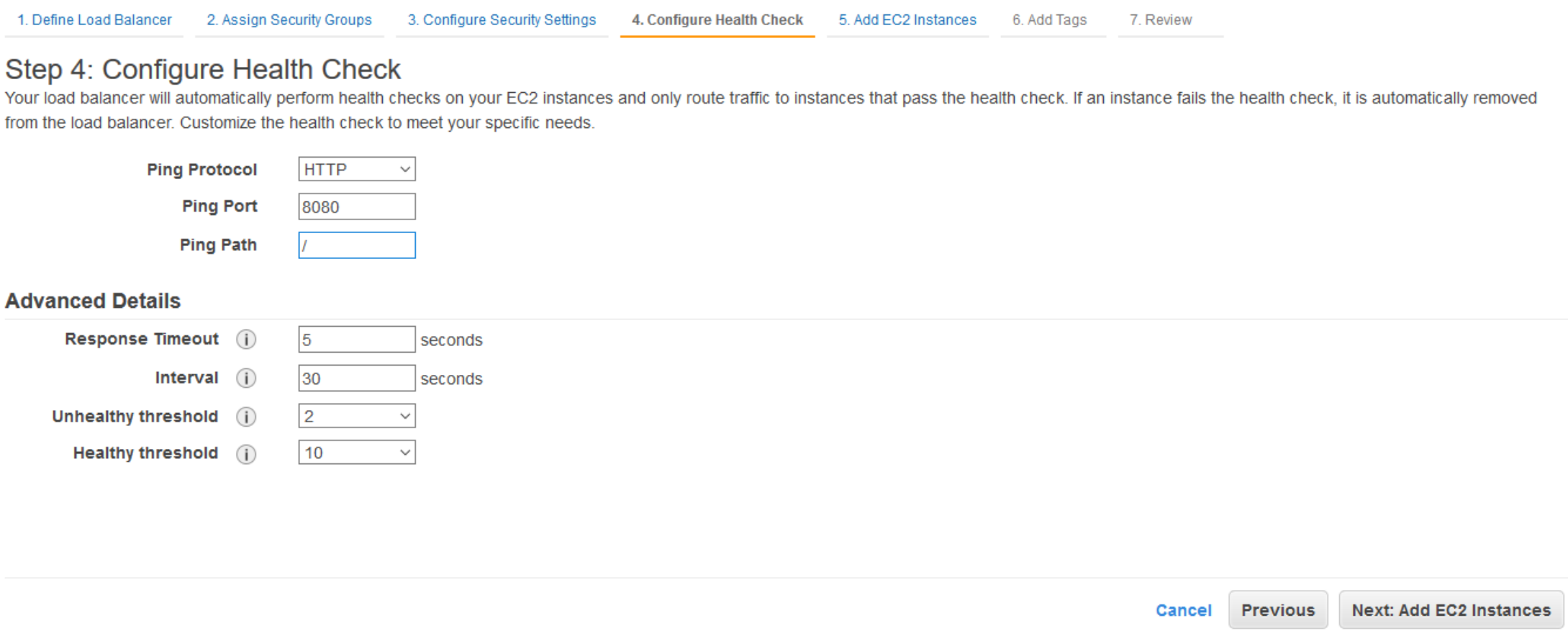


Creating Load Balancer on AWS



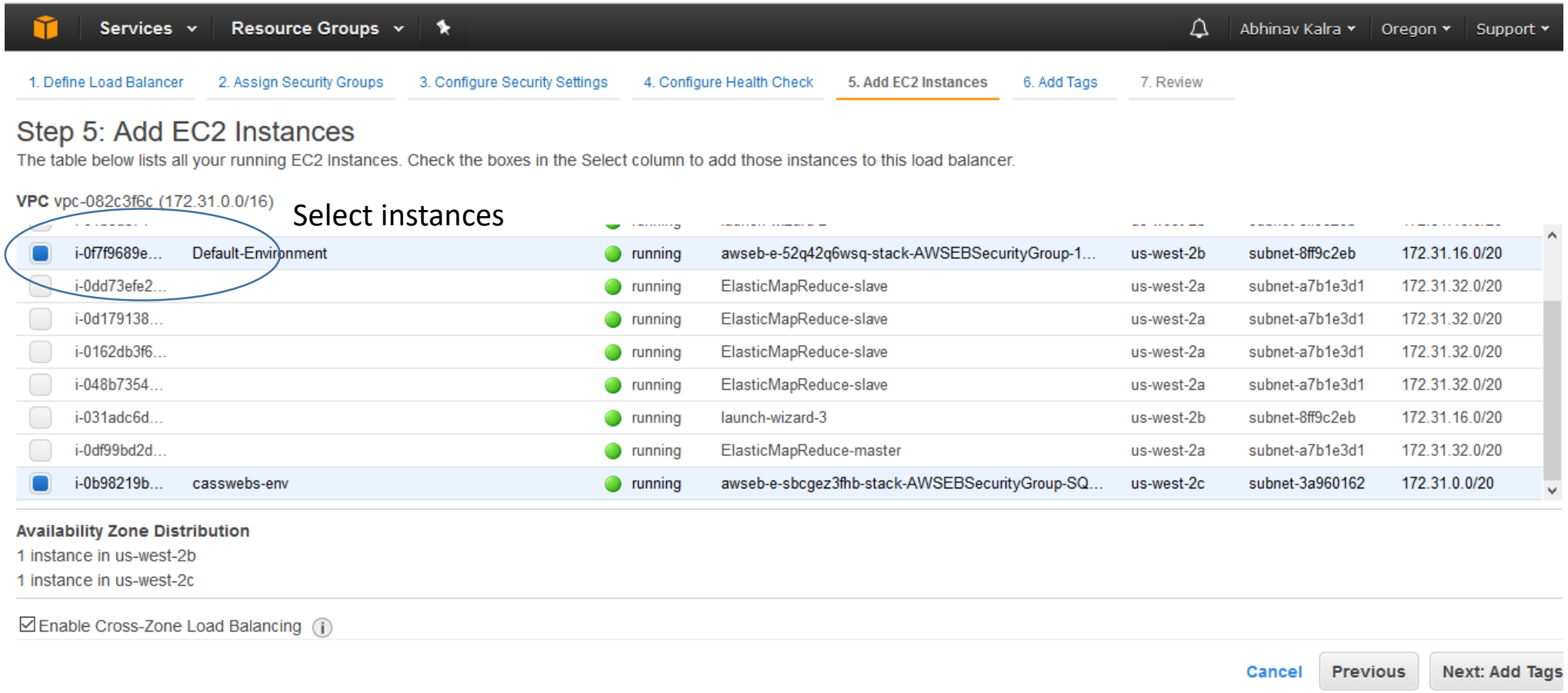
DALHOUSIE UNIVERSITY
Inspiring Minds

Creating Load Balancer on AWS



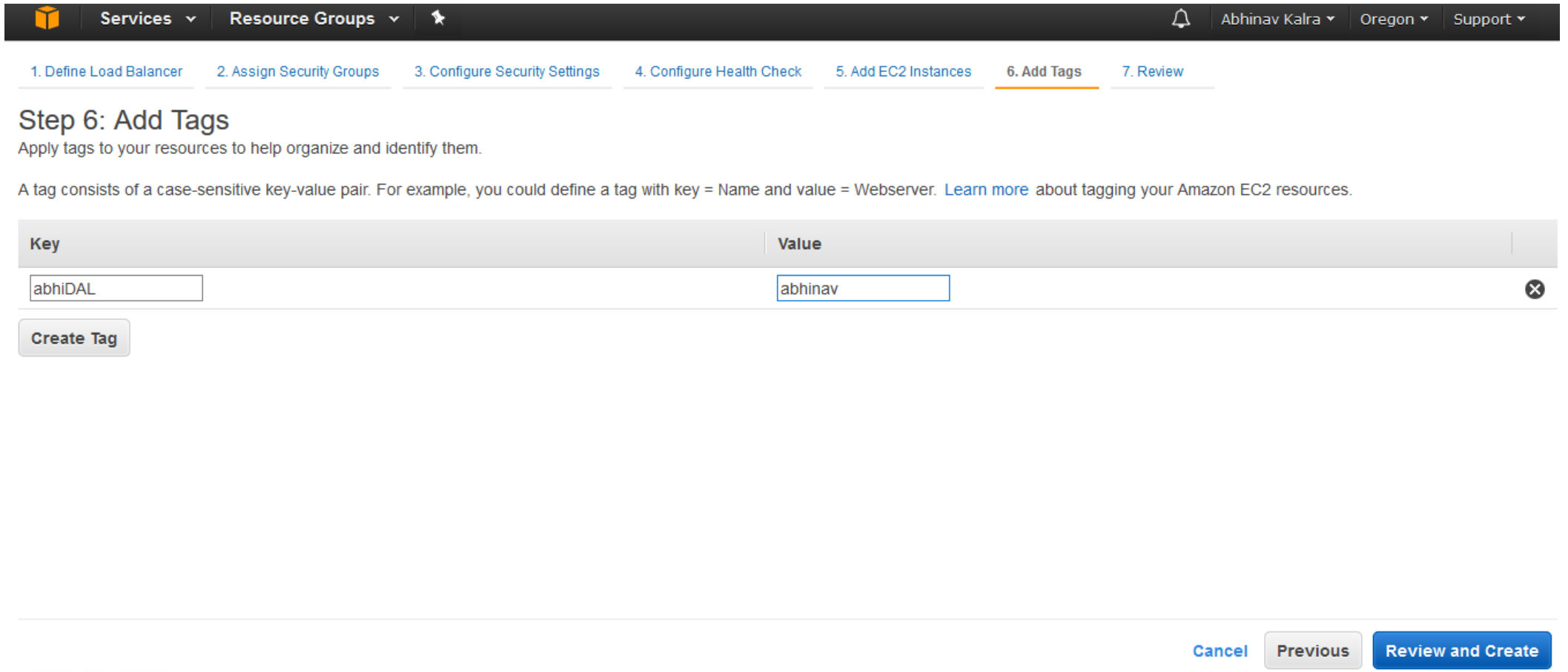
DALHOUSIE UNIVERSITY
Inspiring Minds

Creating Load Balancer on AWS



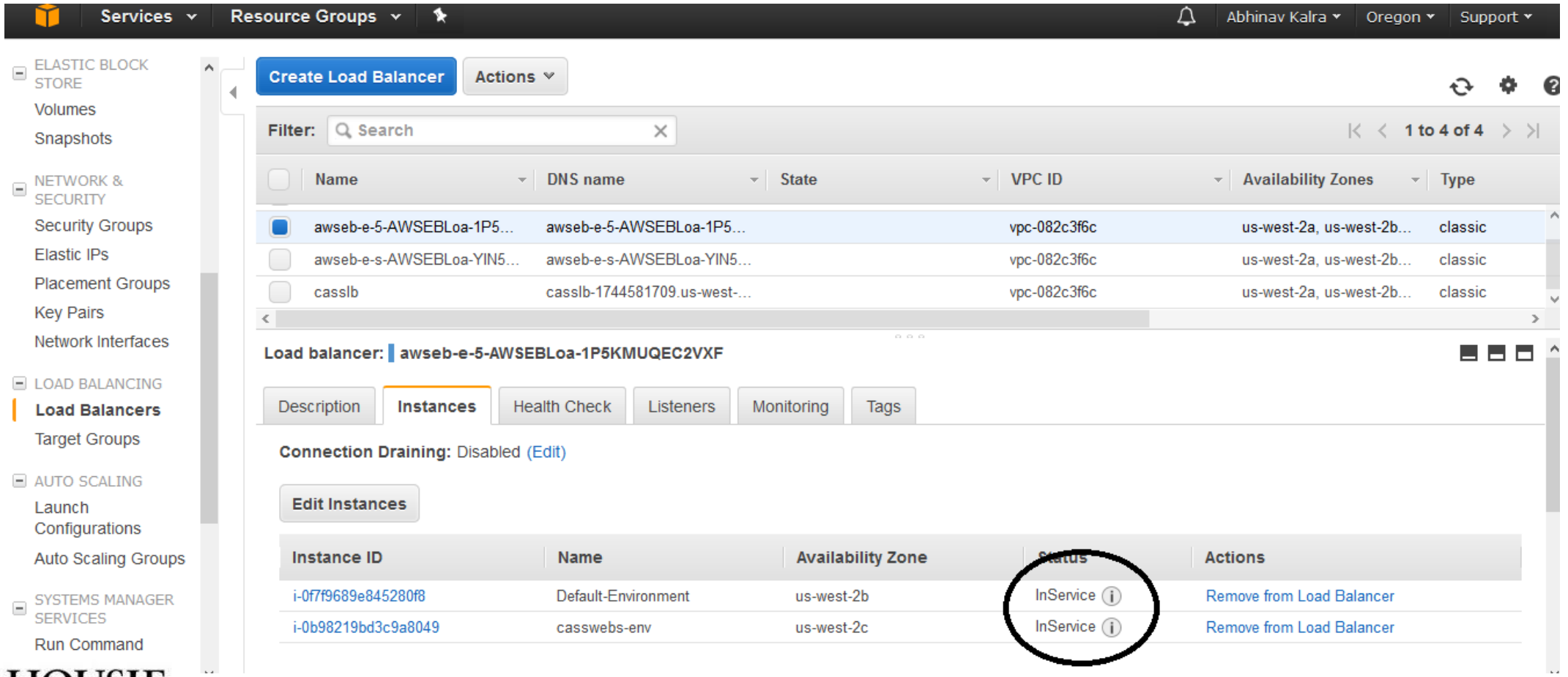
DALHOUSIE UNIVERSITY
Inspiring Minds

Creating Load Balancer on AWS



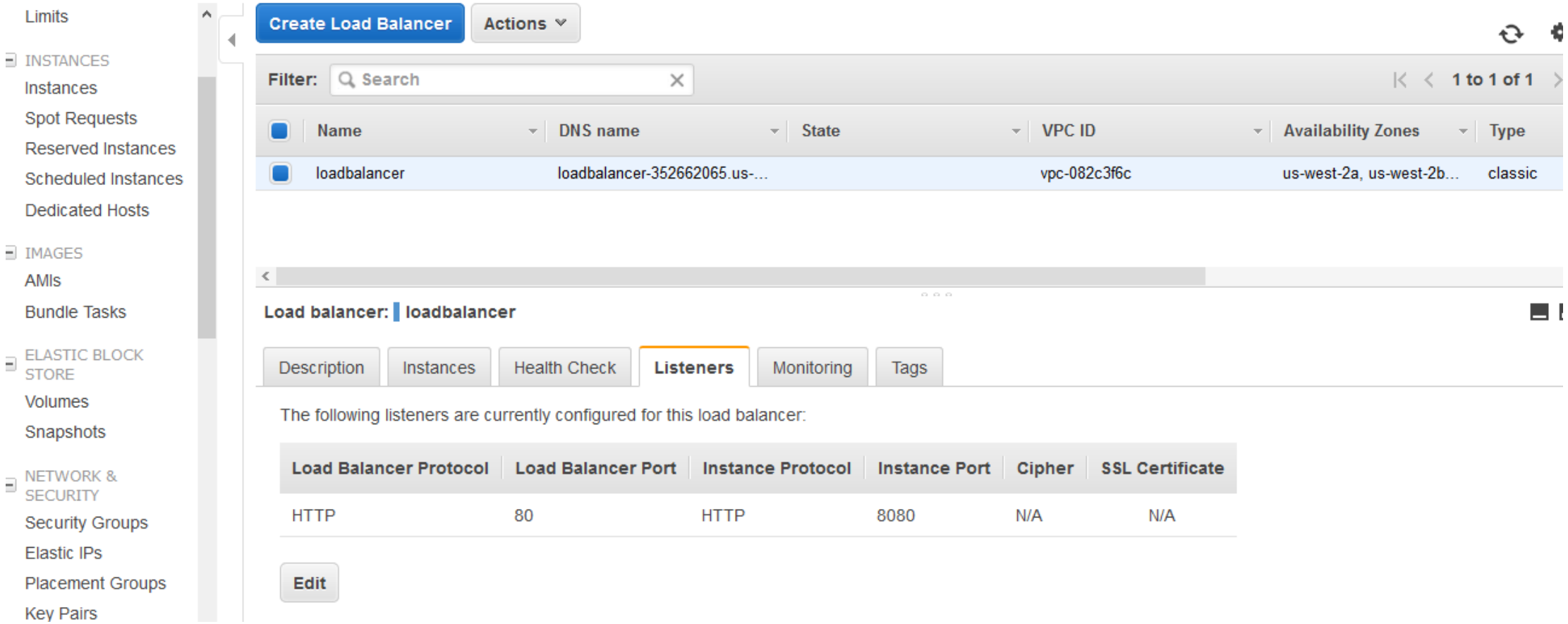
DALHOUSIE UNIVERSITY
Inspiring Minds

Creating Load Balancer on AWS



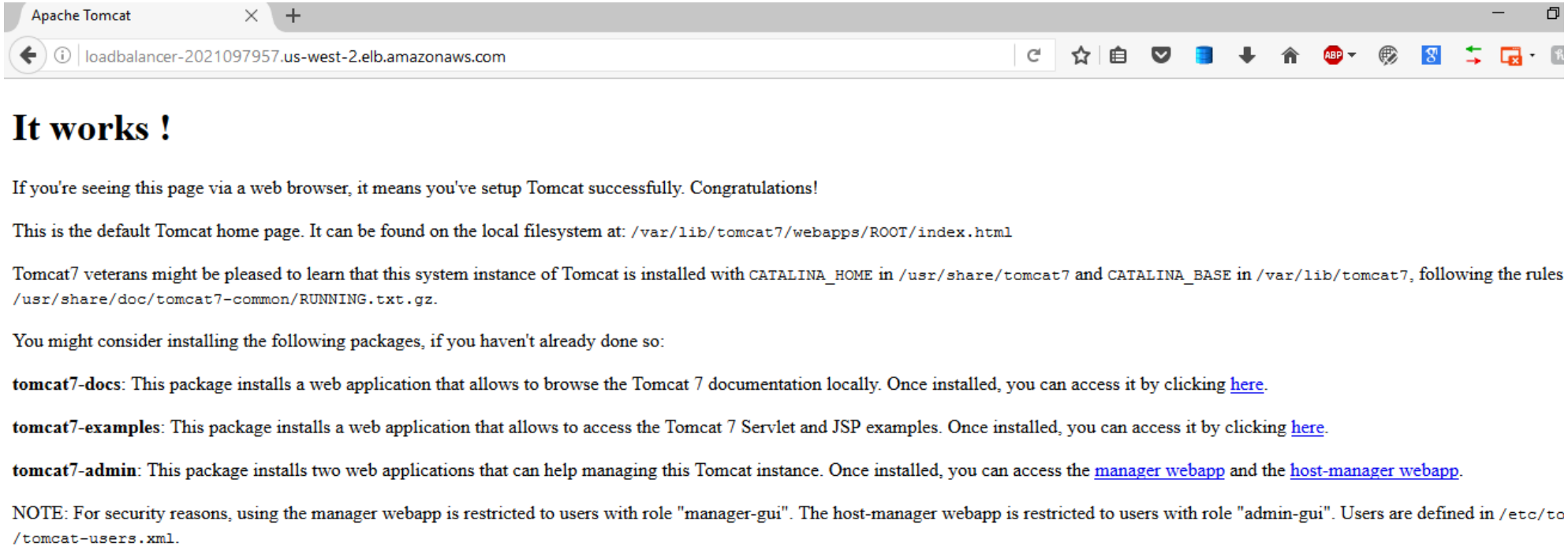
DALHOUSIE UNIVERSITY
Inspiring Minds

Creating Load Balancer on AWS



DALHOUSIE UNIVERSITY
Inspiring Minds

Load Balancer

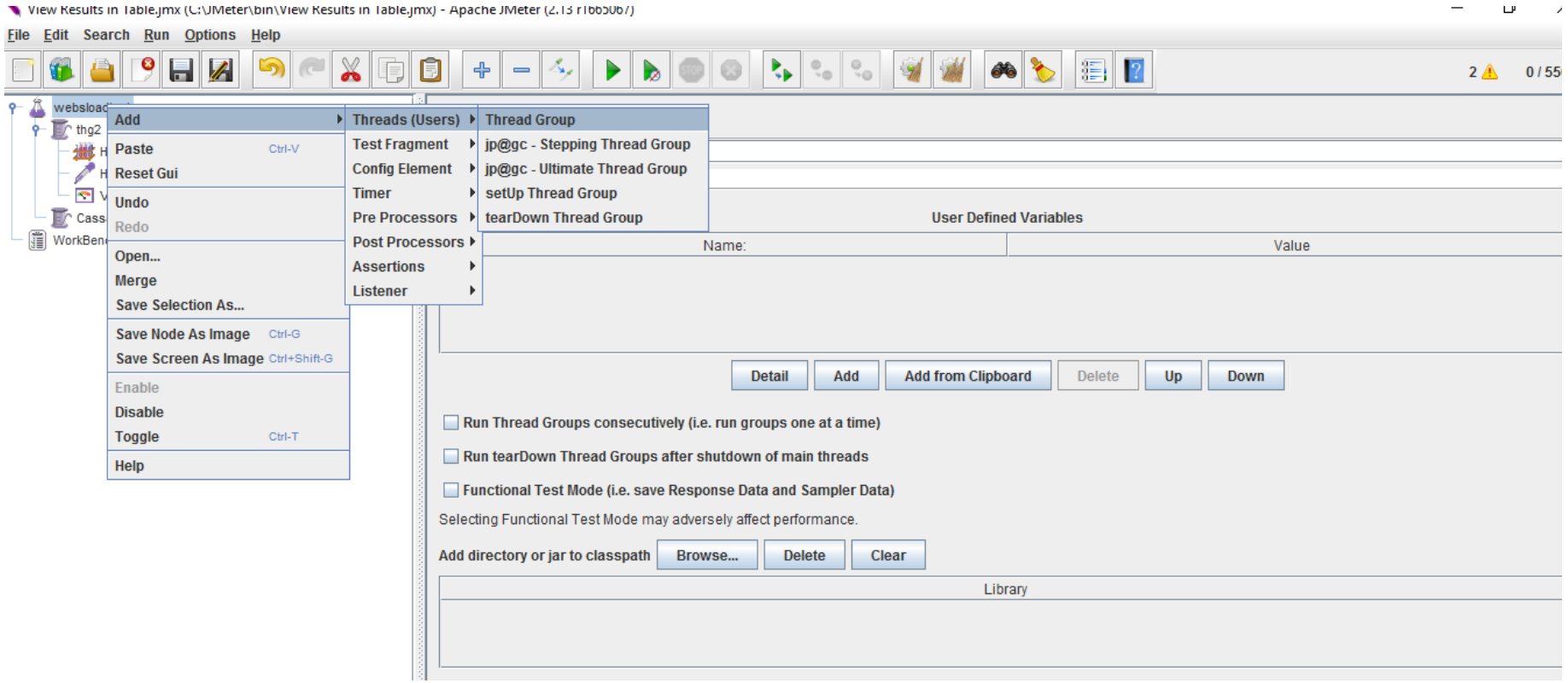


Stress Testing with JMeter

- In this tutorial we are using Apache JMeter
  - Download at : http://jmeter.apache.org/
- To perform stress testing and evaluate load balancing, ensure that your web services are deployed on more than one cloud instance
- These instances will be incrementally added or removed from load balancer while performing stress test.

Running test on JMeter

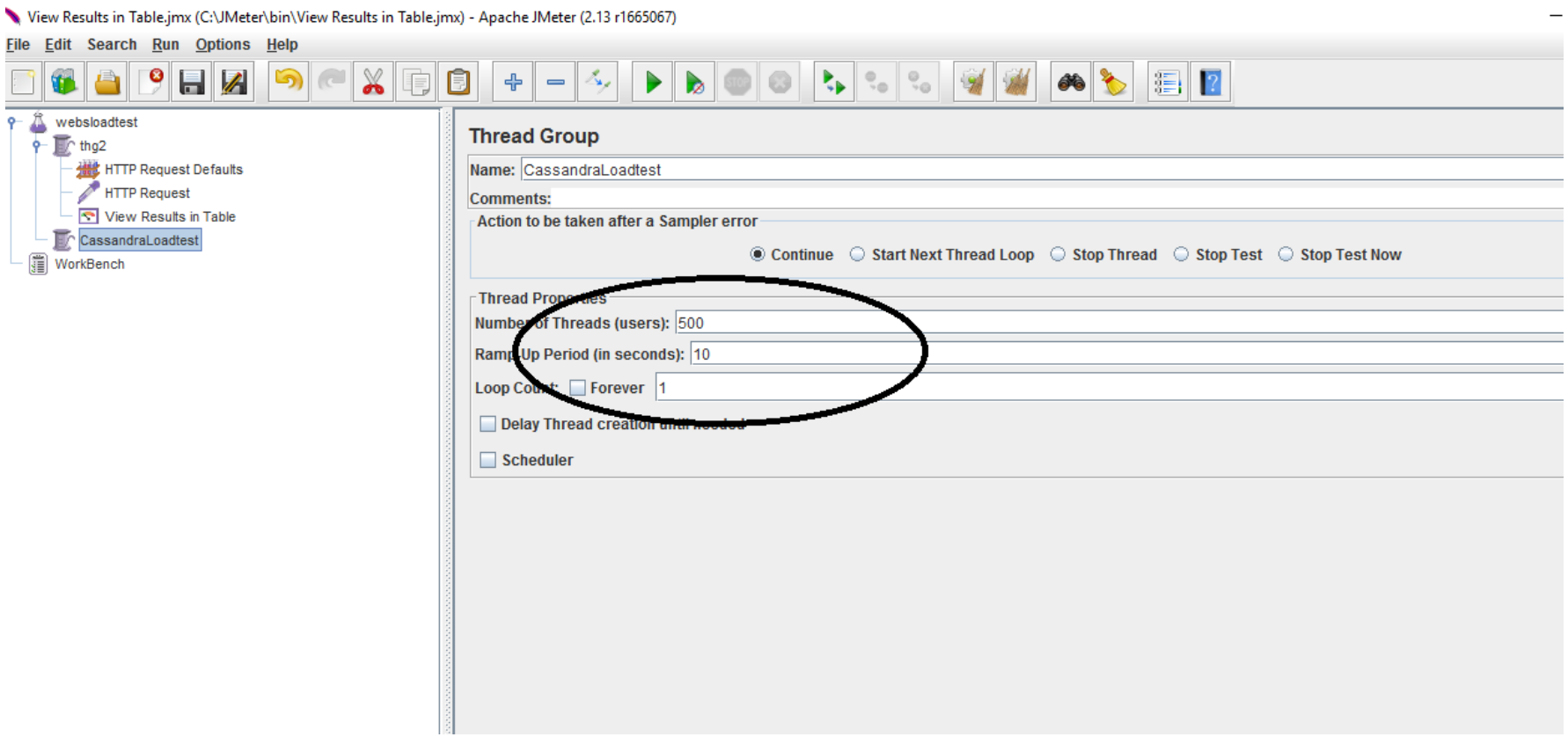
- Create a thread group



DALHOUSIE UNIVERSITY
Inspiring Minds

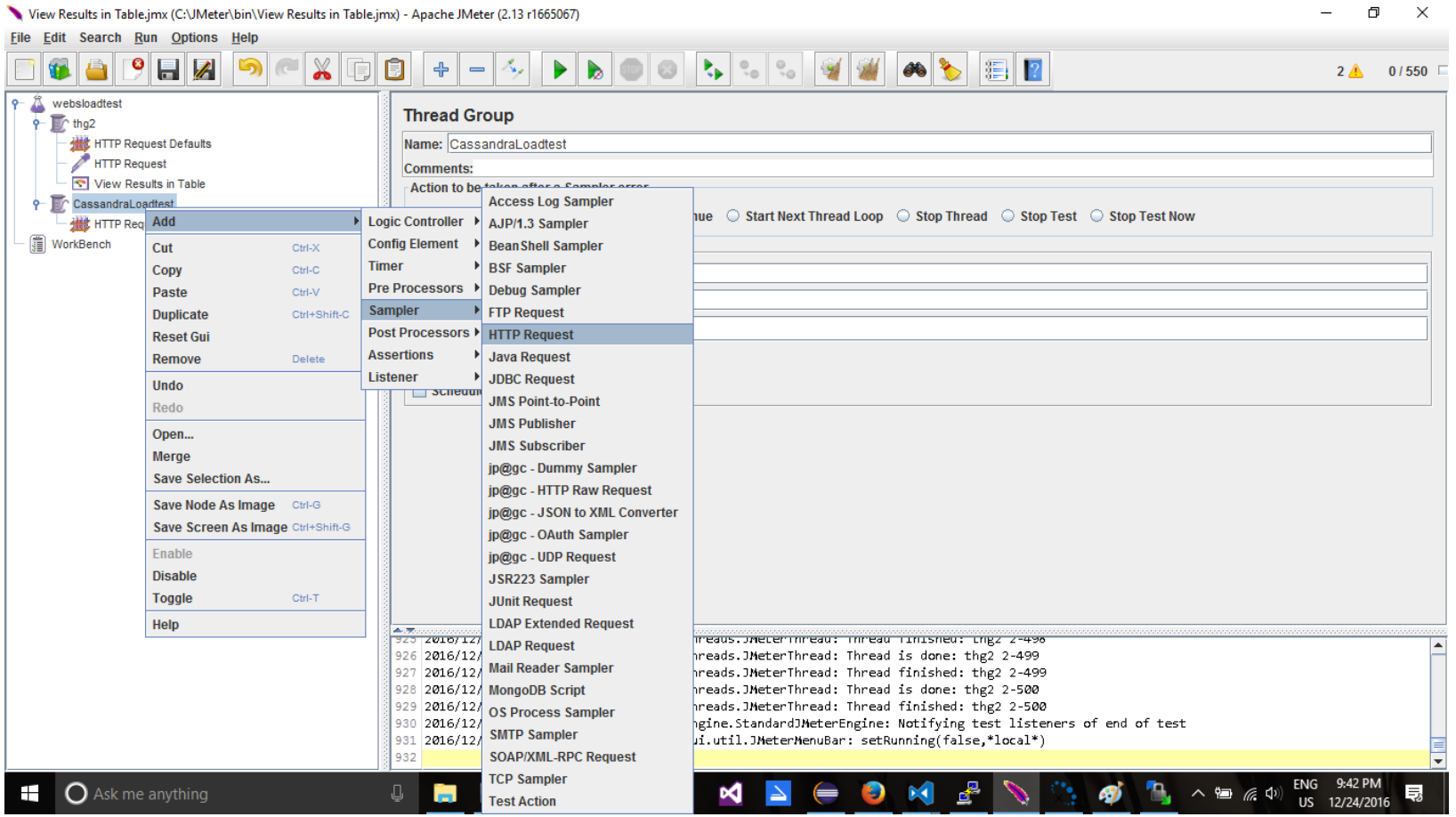
Running test on JMeter

- Specify threads



Running test on JMeter

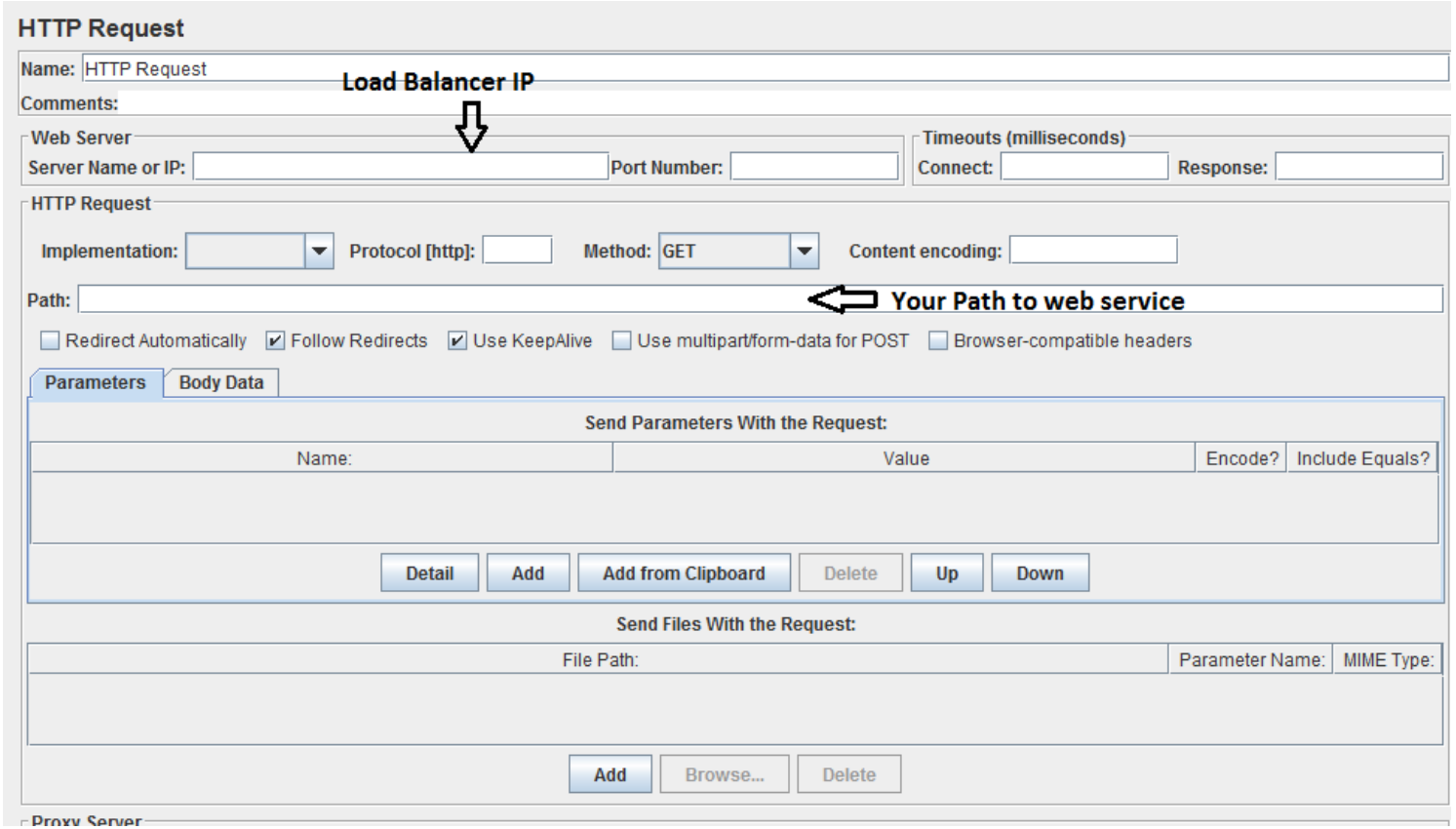
- Add HTTP Request



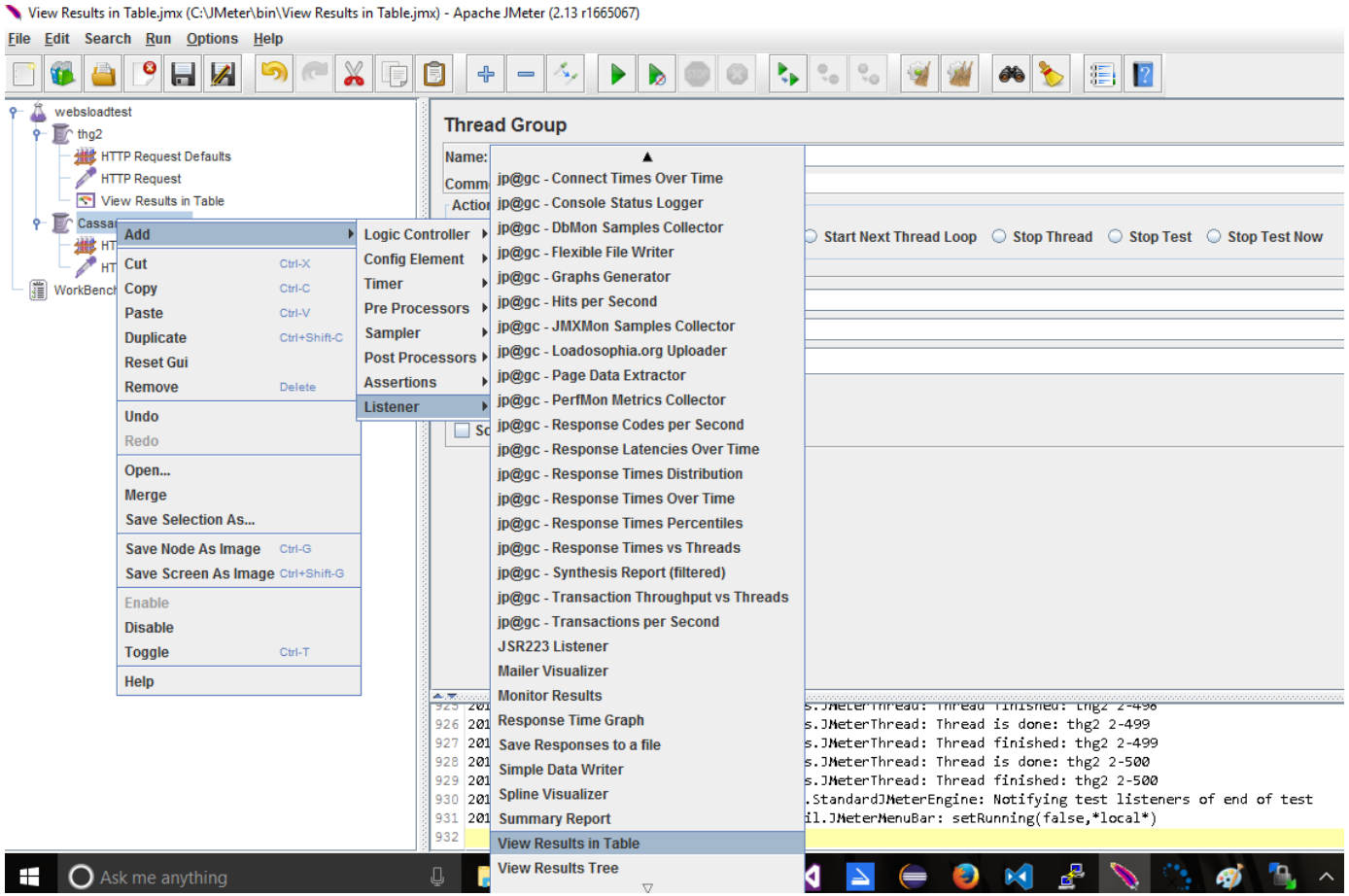
DALHOUSIE UNIVERSITY
Inspiring Minds

Running test on JMeter

- Specify parameters and URI path

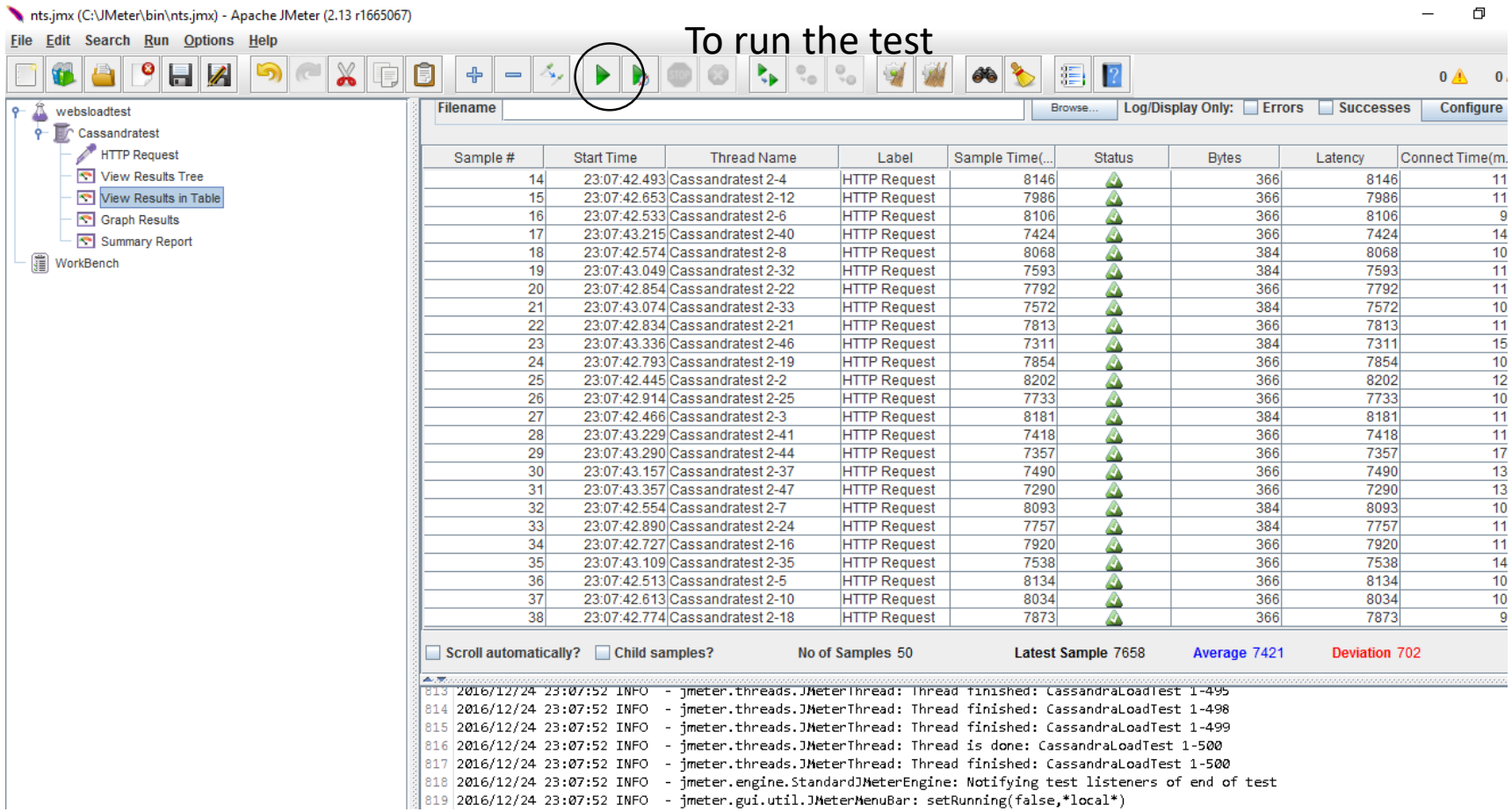


Running test on JMeter



DALHOUSIE UNIVERSITY
Inspiring Minds

Running test on JMeter



DALHOUSIE UNIVERSITY
Inspiring Minds

Assignment Tasks

1. Create a hosted Cassandra cluster
2. Download the dataset
3. Create your schema in Cassandra and upload data
4. Define your application scenario and queries
5. Test your queries by executing them on the server
6. Create a UI client to query data from Cassandra
   - Use Java/.Net/Python
7. Run queries in UI client and return results

Assignment Tasks

8. Create a RESTful web service to query Cassandra DB
9. Use deployed web service to query Cassandra and fetch results.
10. Deploy web service on additional cloud instances for further tasks
11. Create your load balancer and add cloud instances to it
12. Perform stress test with varying values
13. Remove/stop a cloud instance and repeat step 12
14. Evaluate performance and analyze results

Submission Requirements

1. Task description, application scenario and requirements
2. Dataset description
3. Overview of configuration steps with screenshots
4. Brief explanation of application queries
5. CQL syntax of the queries together with execution time
6. Results returned from the query
7. Public DNS and service endpoint URL of web services
8. Test results from Load Balancing task

Submission Requirements

9. Analysis and observations
10. Comments and recommendation on using the software tools
11. Alongside report submit :
    I. Source code for UI and web service
    II. Scripts for Cassandra setup/schema definition etc.
    III. Output from queries
    IV. Any files/ statistics generated from load test tool
    V. README file detailing installation and execution instructions with additional dependencies/requirements (if any)

References

- "A Brief Introduction To Apache Cassandra | Datastax Academy: Free Cassandra Tutorials And Training". *Academy.datastax.com*. N.p., 2016.
- "Cassandra Tutorial". *www.tutorialspoint.com*. N.p., 2016.
- "Java Driver For Apache Cassandra - Home". Docs.datastax.com. N.p., 2016. Web. 14 Dec. 2016.
- "Documentation". Cassandra.apache.org. N.p., 2016.